Хмельницький національний університет
Міністерство освіти і науки України

Кваліфікаційна наукова праця
на правах рукопису

МОЛЧАНОВА МАРИНА ОЛЕКСІЇВНА

УДК 004.8

Д И С Е Р Т А Ц І Я

МЕТОДИ ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ ПРИЙОМІВ ТА ОБ'ЄКТІВ ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

122 Комп'ютерні науки

12 Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

Молчанова Марина Олексіївна

Наукові керівники: Бармак Олександр Володимирович,
доктор технічних наук, професор

Паван Кумар Датт,
доктор філософії,
старший викладач Школи права
Таллінського технічного
університету (Естонія)

Хмельницький – 2025

АНОТАЦІЯ

Молчанова Марина Олексіївна. Методи виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті засобами штучного інтелекту. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії з галузі знань 12 Інформаційні технології за спеціальністю 122 Комп’ютерні науки. – Хмельницький національний університет, Хмельницький, 2025.

Пропаганда, прихована під виглядом звичайних новин, поширюється вже багато десятиліть, однак сучасна цифрова епоха створює сприятливі умови для ще швидшого і більш масового її розповсюдження. Нові методи генерації текстів, які дедалі менше відрізняються від створених людиною, призводять до стрімкого зростання кількості контенту, зокрема й пропагандистського. Це свідчить про необхідність створення автоматизованих методів виявлення пропагандистських маніпуляцій, що сприятиме усвідомленому сприйняттю інформації користувачами та забезпеченню безпеки інформаційного середовища.

На сьогодні ефективними засобами для автоматизованого виявлення та класифікації прийомів і об’єктів пропаганди у текстовому контенті є засоби штучного інтелекту. Однак, попри бурхливий розвиток галузі обробки природної мови, відсутній комплексний інструмент, який би не лише забезпечував виявлення та класифікацію прийомів пропаганди, а й показував на кого і на що вона спрямована. Також проблемою використання моделей глибокого навчання є їх низька інтерпретованість, що породжує недовіру до нейромережевого виявлення пропаганди.

Об’єкт дослідження: процес інтелектуального аналізу текстового контенту для виявлення прийомів та об'єктів пропаганди.

Предмет дослідження: методи та засоби обробки природної мови для виявлення прийомів та об'єктів пропаганди.

Метою дослідження є підвищення точності та якості виявлення прийомів та об’єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень.

У дисертаційній роботі удосконалено метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, який відрізняється від існуючих модифікованою архітектурою нейромережі та обсягом вихідних даних, що дало змогу підвищити точність класифікації.

У дисертаційній роботі розроблено новий метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих використанням доповненої множини семантичних маркерів для виявлення прийомів пропаганди, що дало змогу пояснити отримані результати і підвищити точність та якість виявлення пропаганди.

У дисертаційній роботі також розроблено новий метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих асоціативним групуванням об’єктів пропаганди, що дало змогу покращити результати виявлення об’єктів пропаганди та візуально їх інтерпретувати.

Практичне значення отриманих результатів полягає в доведенні теоретичних результатів дисертаційної роботи до реалізації та у безпосередньому використанні їх у виробничій діяльності підприємств.

Реалізована система комплексного виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті засобами штучного інтелекту. Розроблена інтелектуальна система надає користувачам можливість автоматизованого аналізу текстів, забезпечуючи комплексний результат у вигляді: загальної оцінки прояву пропаганди у тексті; виявлених пропагандистських прийомів із числовими характеристиками сили їх прояву;

оцінки приналежності виявлених об’єктів пропаганди до використаних прийомів. Також інтелектуальна система дозволяє отримувати візуальну інтерпретацію прийнятих рішень, що сприяє підвищенню прозорості та довіри до отриманих результатів.

Результати дисертаційної роботи були впроваджені в освітній процес Хмельницького національного університету при викладанні дисципліни «Методи та системи штучного інтелекту». Апробація та впровадження розробок здійснені в діяльності відділу протидії кіберзлочинам у Хмельницькій області Департаменту кіберполіції Національної поліції України, ГО «ІТ Кластер м. Хмельницького», ПП «Авіві» та ТОВ «Системи для бізнесу 2», а також при виконанні держбюджетної теми Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання» (ДР № 0121U112025).

Зокрема, в діяльності відділу протидії кіберзлочинам у Хмельницькій області Департаменту кіберполіції Національної поліції України знайшли застосування результати дисертаційної роботи: метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання; метод виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень; метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. Ці результати були використані для аналізу наявності пропагандистського контенту в сумнівних матеріалах вебджерел, визначення прийомів та об’єктів пропаганди і візуального пояснення прийнятих рішень. Тестування зазначених методів на різних даних підтвердило високу точність виявлення прийомів і об’єктів пропаганди, зручність та інформаційну достатність запропонованої візуальної інтерпретації аналізу контенту для виявлення пропагандистських проявів.

У діяльності ГО «ІТ Кластер м. Хмельницького» було використано такі результати дисертаційної роботи: метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, що дозволяє визначати рівень наявності політичної пропаганди у повідомленнях; метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, що встановлює присутність пропагандистських маркерів і за ними визначає рівні прояву кожного з використаних прийомів пропаганди у текстових повідомленнях; метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень, що виявляє об'єкти пропаганди та їх зв'язок з використаними прийомами пропаганди. Створене з використанням зазначених результатів програмне забезпечення дозволило автоматизовано виявляти об'єкти та прийоми пропаганди й автоматизовано надавати обґрунтування прийнятих рішень у вигляді їх візуальної інтерпретації, що у випадку використання для систем внутрішнього документообігу та обміну повідомленнями є важливою складовою.

Для апробації та впровадження розробок у ПП «Авіві» було використано: метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання; метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень; метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. Зазначені результати були використані при розробці програмного забезпечення, зокрема соціально орієнтованих сервісів, для автоматизованого детектування наявності пропагандистського контенту, включно з визначенням прийомів та об'єктів пропаганди і візуальним поясненням прийнятих рішень.

Результати роботи було впроваджено у ТОВ «Системи для бізнесу 2», зокрема, метод класифікації текстів за вмістом пропаганди нейромережевими

моделями глибокого навчання, метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. Наведені методи застосовувалися ТОВ «Системи для бізнесу 2» при розробці месенджерів для корпоративного та загального використання у модулях для автоматичного контролю текстового контенту та виявлення соціально шкідливих проявів пропаганди.

Ключові слова: пропаганда, прийоми пропаганди, об’єкти пропаганди, NLP, нейромережі глибокого навчання, BERT, RNN, BiLSTM, NER.

ANNOTATION

Molchanova Maryna. Methods of detecting and classifying techniques and objects of propaganda in text content using artificial intelligence. – Manuscript copyright.

Thesis on competition of scientific degree of Doctor of Philosophy by specialty 122 – Computer Science. – Khmelnytskyi National University, Khmelnytskyi, 2025.

Propaganda, disguised as ordinary news, has been spreading for many decades, but the modern digital era creates favorable conditions for its even faster and more massive distribution. New methods of generating texts, which are increasingly similar to those created by humans, lead to a rapid increase in the amount of content, including propaganda content. This emphasizes the importance of creating automated methods for detecting propaganda manipulations, which will contribute to the conscious perception of information by users and ensuring the security of the information environment.

Today, artificial intelligence tools are effective tools for automated detection and classification of propaganda techniques and objects in text content. However, despite the rapid development of the field of natural language processing, there is no comprehensive tool that would not only provide detection and classification of propaganda techniques, but also show who and what it is aimed at. Another problem with using deep learning models is their low interpretability, which creates distrust in neural network detection of propaganda.

Object of research: the process of intellectual analysis of text content to detect propaganda techniques and objects.

Subject of research: methods and tools of natural language processing to detect propaganda techniques and objects.

The purpose of the research is to increase the accuracy and quality of detection of propaganda techniques and objects by semantic markers in text content using artificial intelligence tools with further explanation of the decisions made.

The dissertation work improved the method of classifying texts by propaganda content using deep learning neural network models, which differs from existing ones in the modified architecture of the neural network and the volume of input data, which made it possible to increase the accuracy of classification.

The dissertation work also developed a new method for detecting propaganda techniques by markers with visual interpretation of decisions made, which differs from existing ones by using an augmented set of semantic markers to detect propaganda techniques, which made it possible to explain the results obtained and increase the accuracy and quality of propaganda detection.

The dissertation work developed a new method for detecting propaganda objects using deep learning neural network models with visual interpretation of decisions made, which differs from existing ones by associative grouping of propaganda objects, which made it possible to improve the results of detecting propaganda objects and visually interpret them.

The practical significance of the results obtained lies in bringing the theoretical results of the dissertation work to implementation and in their direct use in the production activities of enterprises.

A system for comprehensive detection and classification of propaganda techniques and objects in text content using artificial intelligence has been implemented. The developed intellectual system provides users with the ability to perform automated text analysis, providing a comprehensive result in the form of: a general assessment of the manifestation of propaganda in the text; identified propaganda techniques with numerical characteristics of the strength of their manifestation; identified propaganda objects with an assessment of their belonging to the techniques used. The intellectual system also allows for a visual interpretation of the decisions made, which contributes to increasing transparency and trust in the results obtained.

The results of the dissertation work were implemented in the educational process of Khmelnytskyi National University when teaching the discipline "Artificial

Intelligence Methods and Systems". The development was tested and implemented at the Cybercrime Countermeasures Department in Khmelnytskyi Oblast of the Cyberpolice Department of the National Police of Ukraine, NGO "IT Cluster of the city of Khmelnytskyi", Private Enterprise "Avivi" and LLC "Systems for Business 2", as well as in the implementation of the state budget theme of Khmelnytskyi National University "Development of information technology for making human-controlled critical safety decisions based on mental-formal machine learning models" (ДР No. 0121U112025).

In particular, the results of the doctoral thesis were applied in the activities of the Cybercrime Countermeasures Department in Khmelnytskyi Oblast of the Cyberpolice Department of the National Police of Ukraine: a method of classifying texts by propaganda content using deep learning neural network models; a method of identifying propaganda techniques using markers with visual interpretation of decisions made; a method of identifying propaganda objects using deep learning neural network models with visual interpretation of decisions made. These results were used to analyze the presence of propaganda content in suspicious web source materials, identify propaganda techniques and objects, and visually explain decisions made. Testing of these methods on various data confirmed the high accuracy of detecting propaganda techniques and objects, the convenience and information sufficiency of the proposed visual interpretation of content analysis for identifying propaganda manifestations.

In activity of NGO "IT Cluster of Khmelnytskyi" used the following results of the doctoral thesis: a method of classifying texts by propaganda content using deep learning neural network models, which allows determining the level of political propaganda in messages; a method of identifying propaganda techniques by markers with a visual interpretation of decisions made, which establishes the presence of propaganda markers, and based on them determines the levels of manifestation of each of the used political propaganda techniques in text messages; a method for identifying objects of political propaganda using deep learning neural network models

with visual interpretation of decisions made, which identifies objects of political propaganda and their connection with the used methods of political propaganda. The software created using the above results allowed for the automated identification of objects and methods of political propaganda and the automated provision of justification for decisions made in the form of their visual interpretation, which in the case of use for internal document management and messaging systems.

As a test and implementation of developments in the Avivi PE, the following were used: a method for classifying texts by propaganda content using deep learning neural network models; a method for identifying propaganda methods by markers with visual interpretation of decisions made; a method for identifying objects of propaganda using deep learning neural network models with visual interpretation of decisions made. The above results were used in the development of software, in particular socially oriented services, for automated detection of the presence of propaganda content, including the definition of methods and objects of propaganda and visual explanation of the decisions made.

Regarding the implementation in LLC "Systems for Business 2", the developed method of classifying texts by propaganda content using deep learning neural network models, the method of detecting propaganda methods using markers with visual interpretation of the decisions made, the method of detecting propaganda objects using deep learning neural network models with visual interpretation of the decisions made were used. The above methods were used by LLC "Systems for Business 2" when developing messengers for corporate and general use in modules for automatic control of text content and detection of socially harmful manifestations in the form of propaganda.

Key words: propaganda, propaganda techniques, propaganda objects, NLP, deep learning neural networks, BERT, RNN, BiLSTM, NER.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у наукових виданнях, включених до Переліку наукових фахових видань України:

1. Молчанова М.О. Нейромережеве виявлення і класифікація прийомів та об'єктів пропаганди у текстовому контенті. *Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах»*. 2024. № 4. С. 153–161. (https://doi.org/10.31891/2219-9365-2024-80-19).

2. Молчанова М.О. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. *Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки*. 2024. № 5. С. 344–350. (https://doi.org/10.31891/2307-5732-2024-341-5-51).

3. Молчанова М.О., Бармак О.В. Метод нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Computer Science and Applied Mathematics»*. 2024. №2. с. 72-82. (https://doi.org/10.26661/2786-6254-2024-2-08).

4. Молчанова М. Метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки*. 2024. № 6. Т. 1. С. 179–185. (https://doi.org/10.31891/2307-5732-2024-343-6-27).

Публікації, які засвідчують апробацію матеріалів дисертації:

5. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification / V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak. *12th International Conference on Advanced Computer Information Technologies (ACIT 2022)*, 2022, pp. 400–405. URL:

https://doi.org/10.1109/ACIT54803.2022.9913162 (*індексована в наукометричній базі Scopus*).

6. Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble / I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Sobko, O. Zalutska, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3806, pp. 312–324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf (*індексована в наукометричній базі Scopus*).

7. Method for Neural Network Detecting Propaganda Techniques by Markers With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, E. Manziuk, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3790, pp. 158–170. URL: https://ceur-ws.org/Vol-3790/paper14.pdf (*індексована в наукометричній базі Scopus*).

8. Method for Detecting Propaganda Objects Using Deep Learning Neural Network Models With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3933, pp. 38–50. URL: https://ceur-ws.org/Vol-3933/Paper_4.pdf (*індексована в наукометричній базі Scopus*).

9. Method of Semantic Features Estimation for Political Propaganda Techniques Detection Using Transformer Neural Networks / I. Krak, M. Molchanova, V. Didur, O. Sobko, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3917, pp. 286–297. URL: https://ceur-ws.org/Vol-3917/paper56.pdf (*індексована в наукометричній базі Scopus*).

Публікації, які додатково відображають наукові результати дисертації:

10. А. с. № 133374 Україна. Комп’ютерна програма «Інтелектуальна інформаційна система для класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання» / М.О. Молчанова. 2025.

11. А. с. № 133375 Україна. Комп’ютерна програма «Інтелектуальна інформаційна система для виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

12. А. с. № 132910 Україна. Комп’ютерна програма «Інтелектуальна інформаційна система для виявлення об’єктів пропаганди у текстовому контенті нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ГО (NGO) – громадська організація

ІТ – інформаційні технології

ПП (PE) – приватне підприємство

NLP – natural language processing

BERT – bidirectional encoder representations from transformers

RNN – recurrent neural networks

BiLSTM – bidirectional long short-term memory

NER – named entity recognition

МЗС – Міністерство закордонних справ

ООН – Організація об’єднаних націй

ЗМІ – засоби масової інфорації

GPT – generative pre-trained transformer

RoBERTa – robustly optimized BERT pretraining approach

LLM – Large language model

QCRI – Qatar computing research institute

MVPROP – multi-view propaganda detection model

TF-IDF – term frequency-inverse document frequency

SVM – support vector machine

k-NN – k-nearest neighbor

LSTM – Long short-term memory

Bi-GRU – bidirectional version of a GRU

DeBERTa – decoding-enhanced BERT with disentangled attention

IAA – inter-annotator agreement

НМ – нейронна мережа

ML – з англ. machine learning, машинне навчання

LIME – Local Interpretable Model-agnostic Explanations

TP (True Positive) – кількість правильно передбачених позитивних випадків

TN (True Negative) – кількість правильно передбачених негативних випадків

FP (False Positive) – кількість хибно передбачених позитивних випадків

FN (False Negative) – кількість хибно передбачених негативних випадків

ВСТУП

Актуальність теми. Пропагандистський контент, замаскований під звичайні новини, поширюється протягом десятиліть, однак сучасна цифрова епоха додатково створює сприятливі умови для ще швидшого і більш масового його розповсюдження. Розробляються нові методи генерації текстів, які чимраз менше відрізняються від створених людиною, що призводить до стрімкого збільшення обсягів пропагандистських матеріалів.

У сучасному інформаційному суспільстві проблема пропаганди набуває все більшої актуальності, викликаючи потребу в ефективних інструментах для її ідентифікації та аналізу. В умовах миттєвого поширення інформації пропагандистські наративи безперешкодно проникають у масову свідомість впливаючи на громадську думку. Це свідчить про необхідність створення автоматизованих методів виявлення пропагандистських маніпуляцій, що допоможе користувачам усвідомлено споживати інформацію та сприятиме створенню безпечних інформаційних середовищ.

Задача виявлення пропаганди включена в План пріоритетних дій Уряду на 2024 рік [1] як задача «Розбудова міжнародного співробітництва з метою підвищення ефективності протидії інформаційним загрозам та боротьби з дезінформацією і російською пропагандою», покладеною на МЗС України.

Виявлення та класифікація прийомів та об’єктів пропаганди у текстовому контенті засобами штучного інтелекту сприяють досягненню Цілі сталого розвитку ООН [2] №16 (Мир, правосуддя та сильні інститути) шляхом підвищення прозорості інформаційного простору та зміцнення інституційної довіри, а також Цілі сталого розвитку ООН №4 (Якісна освіта) через розвиток медіаграмотності та критичного мислення серед населення, що дозволяє ефективно протидіяти дезінформації.

Сьогодні ефективними засобами для автоматизованого виявлення та класифікації прийомів і об’єктів пропаганди у текстовому контенті є засоби штучного інтелекту [3]. Однак, попри бурхливий розвиток галузі обробки природної мови, комплексного інструменту, який би забезпечував не лише виявлення та класифікації прийомів пропаганди, а ще й показував на кого і на що вона спрямована, не існує. Також проблемою використання моделей глибокого навчання є низька інтерпретованість, що породжує проблему довіри [4] до нейромережевого виявлення пропаганди.

Виявлення прийомів та об’єктів пропаганди належить да задач обробки природної мови, які набули актуальності в дослідженнях у вітчизняних та іноземних учених. Зокрема, дана тема висвітлювалася такими українськими науковцями, як: Висоцька В. [45, 56, 57], Крак Ю. [51], Бармак О. [33, 34], Марченко О. [7, 8, 12, 14], Анісімов А. [8, 13], Ланде [25, 9, 54, 10, 11], а також іноземних Liu X. [21, 29], Ma K. [21, 29], G. Martino [18, 28, 41, 93, 94, 96, 97], F. Alam [94, 96, 98].

Аналіз методів і засобів виявлення прийомів і об'єктів пропаганди показав їх потенціал для вирішення різних завдань і можливість використання в різних контекстах. Однак вони не забезпечують комплексного аналізу взаємозв'язків прийомів і об'єктів пропаганди в текстах, не враховують узагальнень для об'єктів пропаганди та їх альтернативних згадувань у текстах. Детекція пропаганди, заснована лише на пошуку іменованих сутностей, не дає наочності щодо спрямованості пропаганди через застосування прийомів пропаганди. Водночас прийоми пропаганди, виявлені на рівні документа, не відображають об'єктів спрямування пропаганди. Крім того, у випадку виявлення пропаганди за допомогою задачі розпізнавання іменованих сутностей виникає проблема, що об'єкти пропаганди можуть бути подані не власними назвами і не синонімами, а їх семантична близькість не є очевидною.

Таким чином, наразі існує суперечність між необхідністю ефективного виявлення пропаганди та недосконалістю існуючих методів і засобів, які не враховують комплексний аналіз взаємозв'язків прийомів і об'єктів пропаганди. Це особливо важливо на початкових етапах аналізу текстового контенту. Відтак автоматизоване виявлення прийомів та об’єктів пропаганди, що дозволить комплексно аналізувати взаємозв’язки виявлених прийомів і об’єктів, є *актуальною науково-прикладною задачею*, одним із варіантів розв’язання якої є розробка методів і засобів, які сприятимуть підвищенню точності автоматизованого виявлення пропаганди.

Зазначена науково-прикладна задача відповідає предметній області Стандарту вищої освіти України зі спеціальності 122 – Комп’ютерні науки для третього (освітньо-наукового) рівня вищої освіти, зокрема такому об’єкту вивчення та діяльності, як «процеси обробки інформації у комп’ютерних системах».

Зв’язок роботи з науковими програмами, планами, темами. Наведені в дисертації дослідження проводились у межах виконання науково-дослідної роботи за держбюджетною темою Хмельницького національного університету №0121U112025 «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання», у якій виконувалась робота над розробкою архітектури моделей глибокого навчання та дослідження їх поясненості різними методами візуальної аналітики, які були застосовані у дисертаційному дослідженні.

Мета і задачі дослідження. *Об’єкт дослідження* – процес інтелектуального аналізу текстового контенту для виявлення прийомів та об’єктів пропаганди.

Предмет дослідження – методи та засоби обробки природної мови для виявлення прийомів та об'єктів пропаганди.

Метою дисертаційного дослідження є підвищення точності та якості виявлення прийомів та об’єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень.

Для досягнення поставленої мети необхідно вирішити відповідні *задачі*:

1. Провести аналіз методів, засобів та технологій для автоматизованого виявлення пропаганди у текстовому контенті.
2. Удосконалити метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання.
3. Розробити метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень.
4. Розробити метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.
5. Розробити інтелектуальну інформаційну систему для валідації запропонованих методів і провести експериментальні дослідження.

Методи дослідження. Для розв’язання поставлених задач використовуються методи математичної статистики та теорії вірогідності, методи машинного навчання, методи обробки природної мови, методи емпіричного дослідження.

Наукова новизна отриманих результатів полягає в досягненні таких наукових результатів:

1. Удосконалено метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, який відрізняється від існуючих модифікованою архітектурою нейромережі та обсягом вихідних даних, що дало змогу підвищити точність класифікації.

2. Розроблено новий метод виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих використанням доповненої множини семантичних маркерів для виявлення прийомів пропаганди, що дало змогу пояснити отримані результати і підвищити точність та якість виявлення пропаганди.

3. Розроблено новий метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих асоціативним групуванням об’єктів пропаганди, що дало змогу покращити результати виявлення об’єктів пропаганди та візуально їх інтерпретувати.

Практичне значення отриманих результатів. Практичне значення отриманих результатів полягає в доведенні теоретичних результатів дисертаційної роботи до реалізації та у безпосередньому використанні їх у виробничій діяльності підприємств.

Реалізована система комплексного виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті засобами штучного інтелекту. Розроблена інтелектуальна система надає користувачам можливість автоматизованого аналізу текстів, забезпечуючи комплексний результат у вигляді: загальної оцінки прояву пропаганди у тексті; виявлених пропагандистських прийомів із числовими характеристиками сили їх прояву; оцінки приналежності виявлених об’єктів пропаганди до використаних прийомів. Також інтелектуальна система дозволяє отримувати візуальну інтерпретацію прийнятих рішень, що сприяє підвищенню прозорості та довіри до отриманих результатів.

Результати дисертаційної роботи впроваджено у (Додаток Б): Відділі протидії кіберзлочинам в Хмельницькій області Департаменту кіберполіції Національної поліції України (довідка про впровадження від 19.03.2025 р.);

ГО «ІТ Кластер м. Хмельницького» (довідка про впровадження від 10.12.2024 р.); ПП «Авіві» (довідка про впровадження від 25.10.2024 р.); ТОВ «Системи для бізнесу 2» (довідка про впровадження від 17.01.2025 р.); навчальному процесі Хмельницького національного університету (акт впровадження від 27.11.2024 р.); при виконанні держбюджетної теми Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання» (ДР № 0121U112025).

Особистий внесок здобувача та внесок інших співавторів у спільних публікаціях. Усі наукові результати дисертаційного дослідження отримані автором особисто. Список опублікованих праць за темою дисертації представлено в списку використаних джерел [42, 43, 107, 117, 120, 123, 124, 132, 134, 149, 152, 155]. У спільних публікаціях автору належать такі результати: метод нейромережевого виявлення прийомів пропаганди із застосуванням візуальної аналітики для інтерпретованості результатів моделей глибокого навчання [43]; метод виявлення політичної пропаганди в інтернет-контенті нейромережевими засобами обробки природної мови та його формальне представлення [117]; метод виявлення об’єктів пропаганди з візуальними представленнями [123]; огляд можливих для використання моделей української мови, засоби попередньої обробки текстових даних для поліпшення процесу класифікації, робота над методологією [124]; метод нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень [132]; метод оцінки семантичних ознак для виявлення прийомів політичної пропаганди за допомогою нейронних мереж архітектури трансформер [134].

Особистий внесок інших співавторів у спільних публікаціях: у публікації [43] Ю. Крак спільно з О. Бармаком виконували концептуалізацію дослідження

та керівництво проєктом, О. Мазурець виконував рецензування і коригування рукопису та адміністрування проєкту, Е. Манзюк та О. Залуцька опрацьовували результати експериментів; у публікації [117] Ю. Крак спільно з О. Бармаком виконували концептуалізацію дослідження та керівництво проєктом, О. Мазурець виконував рецензування і коригування рукопису, В. Дідур створював програмне забезпечення, О. Залуцька та О. Собко опрацьовували результати експериментів; у публікації [123] Ю. Крак спільно з О. Бармаком виконували концептуалізацію дослідження та керівництво проєктом, О. Мазурець керував постановкою експериментів, виконував рецензування і коригування рукопису, О. Залуцька виконувала розробку програмного забезпечення для експериментів; у публікації [124] Ю. Крак виконував концептуалізацію дослідження, проводив обговорення результатів дослідження, О. Бармак – рецензування та коригування рукопису, керівництво проєктом, О. Мазурець виконував концептуалізацію дослідження, огляд відомих методів та рішень, підготовку чернетки рукопису, адміністрування проєкту, О. Ковальчук спільно з В. Слободзяном розроблювали програмне забепечення для апробації запропонованого підходу, автор М. Молчанова спільно з О. Собко працювали над методологією дослідження, а також опрацьовували результати дослідження та підготовку чернетки рукопису; у публікації [132] О. Бармак виконував концептуалізацію дослідження та керівництво проєктом; у публікації [134] Ю. Крак та О. Бармак проводили концептуалізацію дослідження, О. Мазурець – рецензування і коригування рукопису, В. Дідур створював програмне забезпечення, а О. Собко опрацьовувала результати експериментів.

Апробація матеріалів дисертації. Основні результати дисертаційного дослідження доповідались та обговорювались на міжнародних і всеукраїнських науково-технічних та науково-практичних конференціях і семінарах, а саме: 12th International Conference on Advanced Computer Information Technologies

«ACIT’2022)» (September 26–28, 2022), 14th International Scientific and Practical Programming Conference «UkrPROG 2024» (May 14–15, 2024, Kyiv, Ukraine), 12th International Conference Information Control Systems & Technologies «ICST 2024» (September 23–25, 2024, Odesa, Ukraine), Information Technology and Implementation Workshop: Intelligent Systems and Security «IT&I-WS 2024: ISS» (November 20–21, 2024, Kyiv, Ukraine), 7th Workshop for Young Scientists in Computer Science & Software Engineering «CS&SE@SW 2024» (December 27, 2024, Kryvyi Rih, Ukraine).

Публікації. Основні результати дисертації опубліковані у 12 наукових працях [42, 43, 107, 117, 120, 123, 124, 132, 134, 149, 152, 155], Додаток А), зокрема: 4 статті у фахових наукових журналах України, включених на дату опублікування до переліку наукових фахових видань України категорії Б; 5 публікацій, які засвідчують апробацію матеріалів дисертації (публікації, що індексуються в наукометричній базі Scopus); 3 авторських свідоцтва.

Структура та обсяг дисертації. Дисертація складається з анотації, змісту, переліку умовних скорочень, вступу, чотирьох розділів, висновків, списку використаних джерел із 156 найменувань на 23 сторінках і 4 додатків. Загальний обсяг дисертаційної роботи становить 179 сторінок друкованого тексту, із них 138 сторінок основного тексту. Дисертація містить 43 рисунки і 15 таблиць.

РОЗДІЛ 1.
АНАЛІЗ МЕТОДІВ, ЗАСОБІВ ТА ТЕХНОЛОГІЙ ДЛЯ АВТОМАТИЗОВАНОГО ВИЯВЛЕННЯ ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ

1.1. Актуальність задачі виявлення та класифікації пропаганди

Про актуальність задачі виявлення пропаганди свідчить той факт, що Уряд України включив до Плану пріоритетних дій на 2024 рік задачу «Розбудова міжнародного співробітництва з метою підвищення ефективності протидії інформаційним загрозам та боротьби з дезінформацією і російською пропагандою» [1]. Водночас використання ШІ для виявлення та класифікації пропаганди сприяє досягненню Цілей сталого розвитку ООН №16 (Мир, правосуддя та сильні інститути) та №4 (Якісна освіта) [2], зміцнюючи довіру до інституцій та підвищуючи рівень медіаграмотності.

Попри розвиток NLP, комплексного інструменту, що не лише визначає прийоми пропаганди, а й показує її об'єкти та ціль, досі немає. Крім того, низька інтерпретованість глибоких нейромереж залишається викликом для довіри до таких технологій [3, 4].

Пропаганда є невід’ємним складником інформаційних маніпуляцій і включає різноманітні форми, методи і засоби впливу на людей з метою зміни їхніх психологічних характеристик у бажаному напрямі [5], тому її своєчасне виявлення є актуальною задачею інформаційних технологій. Такі маніпуляції часто використовується для зміни психологічного клімату в суспільстві, мобілізації підтримки або дискредитації опонентів.

Значний вплив масмедіа на формування громадської думки спонукає проведенню наукових досліджень пропаганди та маніпулятивного мовного

впливу у ЗМІ, викликає необхідність аналізувати комунікативні чинники в аспекті інформаційної безпеки [6].

Через зростання споживання текстового інтернет-контенту збільшується загроза здійснення пропагандистського деструктивного маніпулятивного впливу політичних медіа. З іншого боку, також розвиваються засоби генеративного штучного інтелекту, проводяться дослідження у сфері пошуку перефразувань [7] та їх генерації [8], виділення прямої мови [9] та виділення подій [10, 11], отримання ефективних прихованих ознак слів і аналізу їх зв'язків у реченні [12], стиснення текстів природною мовою [13], що породжує ще більше можливостей для автоматизованого створення контенту [14] та розповсюдження маніпулятивних впливів. Пропаганда, яка розповсюджується в мережі «Інтернет» представляє масштабну загрозу для національної безпеки країни, несвоєчасне вирішення якої може призвести до руйнівних наслідків [15].

У соціологічному енциклопедичному словнику термін «пропагандистський допис» розглядається у декількох варіантах: 1) поширення в масах ідеології та політики певних класів, партій, держав; 2) засіб маніпуляції масовою свідомістю [16].

Елементи процесу пропаганди містять: суб'єкт, зміст, форми і методи, а також засоби або канали передачі інформації. Суб'єктом пропаганди є соціальна група, яка прагне впливати на аудиторію. Зміст пропаганди визначається соціальними інтересами її суб'єкта та його відношенням до інтересів суспільства загалом. Форми і прийоми пропаганди вибираються залежно від цілей та аудиторії, на яку має здійснюватись вплив. Засоби передачі інформації містять друковані видання, радіо, телебачення, а також лекційну систему тощо. Об’єктом пропаганди є аудиторія або соціальні групи, які є метою впливу. Соціальні інтереси суб'єкта пропаганди впливають на її зміст, вибір форм, методів і засобів передачі інформації [17].

Виявлення пропаганди [18, 19] та дезінформації [20] за допомогою NLP у тексті є складною задачею через тонкі методи маніпулювання та контекстуальні залежності. Щоб вирішити цю проблему, в межах експериментального набору даних SemEval-2020 task 11, який містить статті новин [21, 22] , позначені 14 пропагандистськими прийомами, було виконано ряд досліджень. Автори [23, 24] досліджували ефективність сучасних великих мовних моделей [25], таких як GPT-3 [26], GPT-3.5-Turbo, GPT-4, для виявлення прийомів пропаганди. Використовується п’ять варіантів GPT-3 [27] і GPT-4, які містять різні стратегії оперативного проєктування та тонкого налаштування в різних моделях. Ефективність моделей визначалась за оцінкою таких показників, як F1-score, Precision і Recall та порівнянням результатів з поточним сучасним підходом за допомогою RoBERTa. Однак було отримано результати F1-score 60.2%. Використовуючи технологію, що лежить в основі ChatGPT, дослідники аналізували тексти для визначення присутності різних прийомів пропаганди [28]. Ними було розроблено ретельно уточнений запит, який поєднується зі статтями з мережі Russia Today (RT), щоб визначити наявність пропагандистських методик. Дослідження показало, що технологія LLM може давати розумні висновки про пропаганду, хоча точність виявлення складає всього 25.12%, однак вона демонструє потенціал як інструмент для виявлення пропаганди для кінцевих користувачів – таких, як медіаспоживачі та журналісти.

Авторами [29] помічено, що існуючі методи виявлення пропаганди насамперед зосереджені на виявленні мовних особливостей її змісту. Однак ці методи зазвичай пропускають інформацію, представлену в зовнішньому новинному середовищі, з якого виникли та поширилися пропагандистські новини. Це середовище новин відображає думки основних ЗМІ та увагу громадськості і містить мовні характеристики непропагандистських новин.

Тому в роботі запропоновано мультиінформаційну інтеграційну мережу на основі графів із зовнішнім середовищем новин для виявлення пропаганди.

У дослідженні [30] аналізується, як ЗМІ вплинули та відобразили громадську думку протягом першого місяця війни за допомогою статей і каналів новин у Telegram українською, російською, румунською, французькою та англійською мовами. Запропоновано порівняти два методи багатомовної автоматизованої ідентифікації прокремлівської пропаганди, заснованих на трансформерах і лінгвістичних ознаках. Проаналізовано переваги та недоліки обох методів, їх адаптивність до нових жанрів і мов, а також етичні міркування використання методів для модерації вмісту.

Однак, як було зазначено авторами [31], румунська та українська мови є поки малодослідженими з точки зору інструментів NLP [32] порівняно з англійською, тому є потреба у подальших дослідженнях у напрямі виявлення пропаганди в текстовому інтернет-контенті нейромережевими засобами обробки природної мови.

Досліджено основні методи аналізу газетних текстів для виявлення маніпулятивних технологій, що допомагає застерегти від дезінформації [33, 34] та пропаганди [35]. Представлено новий набір еталонних даних чеською мовою для навчання та оцінки сучасних і майбутніх методів розпізнавання 18 маніпулятивних прийомів, зокрема, нагнітання страху, релятивізація, навішування ярликів. Показано, що поєднання контент-аналізу із запропонованим стильовим аналізом підвищує точність виявлення 15-ти з 17-ти оцінених маніпулятивних прийомів від 0.05% до 1.46%. Метод перевірено на пропагандистській базі QCRI [36]. Подальші дослідження будуть зосереджені на додаванні нових стилометричних характеристик, вдосконаленні існуючих методів та використанні методів доповнення даних для боротьби з дисбалансом етикеток. Також планується перейти до дрібнозернистої класифікації на рівні проміжків часу, а не на рівні документа в цілому.

У ще одному дослідженні представлено багатомовний набір даних про пропаганду та проведено експеримент для дослідження маркерів, за якими людські анотатори та алгоритми класифікації відрізняють пропагандистські статті від непропагандистських на певну тему [37]. Показано, що перебільшення, зменшення описовості та відсутність адекватних джерел часто зустрічаються у пропагандистській пресі. Аналізатор VAGO [38] підтвердив, що використання невизначених маркерів значно корелює з цими особливостями. Виявлено, що моделі машинного навчання ефективні для виявлення пропаганди на певну тему, але потребують покращення щодо пояснюваності та узагальнення на інші теми. Подальші роботи зосередяться на вдосконаленні аналізу, розробці багатомовних моделей та покращенні інструментів пояснюваності. Також планується введення нових міток для уточнення анотацій та ідентифікації більшої кількості стилістичних особливостей.

Застосування моделі MVPROP, що використовує багатовимірні контекстні вбудовування, дозволяє покращити точність виявлення пропаганди. Експерименти показали, що модель може бути перенесена на новинні статті [39]. Для тестування представлено TWEETSPIN [40] – набір даних із твітами, що містить слабкі анотації тонких пропагандистських прийомів, і модель MVPROP для їх виявлення. TWEETSPIN включає лише ідентифікатори твітів, що відповідає умовам використання Twitter, і містить потенційно образливі та ворожі висловлювання. Основним обмеженням є слабкі анотації через великий масштаб даних. У майбутньому планується дослідження виявлення пропаганди на рівні окремих фрагментів.

Як підтверджено у роботах вище, пропаганда характеризується прийомами, за які відповідають певні маркери, що притаманні цим прийомам. У роботі увага зосереджена на виявленні 17-ти відомих прийомів пропаганди, детально описаних в [41], а саме: «Апеляція до авторитету» («Appeal to Authority»), «Помилка хибної дихотомії» («Black and White Fallacy»), «Сумнів»

(«Doubt»), «Причинно-надмірна спрощеність» («Causal Oversimplification»), «Перебільшення» («Exaggeration»), «Розмахування прапором» («Flag-Waving»), «Заряджена мова» («Loaded Language»), «Навішування ярликів» («Labeling»), «Мінімізація» («Minimisation»), «Обзивання» («Name Calling»), «Зведення до Гітлера» («Reductio ad Hitlerum»), «Червоний оселедець» («Red Herring»), «Повторення» («Repetition»), «Гасла» («Slogans»), «Ватабоутізм» («Whataboutism»), «Кліше, що припиняють думання» («Thought Terminating Cliches»), «Апеляція до страху» («Appeal to Fear-Prejudice»).

Різноманітні дослідження підтверджують, що реалізація існуючих методів виявлення пропаганди в текстовому контенті стикається з труднощами, пов'язаними з тонкими методами маніпуляції та контекстуальними залежностями. Застосування сучасних мовних моделей, таких як GPT-3 та GPT-4, BERT-архітектур, демонструє потенціал для розпізнавання пропагандистських прийомів, проте все ще потребує вдосконалення щодо точності та пояснюваності результатів.

З огляду сучасних можливостей застосування штучного інтелекту до виявлення та класифікації прийомів пропаганди визначено актуальним створення нових методів та засобів саме для комплексного аналізу пропагандистського контенту, що може бути реалізовано шляхом виявлення маркерів, прийомів та об’єктів пропаганди як взаємопов’язаних складових моделей пропагандистських впливів.

1.2. Аналіз напрямів виявлення і класифікації прийомів пропаганди

Кінцевою метою застосування засобів та методів штучного інтелекту до виявлення та класифікації прийомів пропаганди є створення автоматизованих інструментів, які допомагають експертам аналізувати новинну екосистему, виявляти спроби маніпуляції та досліджувати, як події, глобальні проблеми та

політики висвітлюються в медіасфері різних країн і різними мовами. Напрям виявлення пропаганди визначає, на якому рівні аналізуються текстові дані для виявлення прийомів пропаганди. Є два основні напрями виявлення пропаганди: напрям на основі виявлення іменованих об’єктів [42] та напрям на основі класифікації текстів [43].

При виявленні пропаганди як завдання виявлення іменованих сутностей виникає проблема, що фрагменти тексту, які містять пропаганду, є довшими за NER (наприклад, назви осіб або місць) і можуть включати десятки слів.

У [44] досліджується величина довжини діапазону, що впливає на розпізнавання пропаганди, демонструючи, що складність завдання справді зростає зі збільшенням довжини діапазону. Систематично оцінюються кілька поширених підходів до завдання, наскільки добре вони відновлюють розподіл довжини справжніх проміжків. Також пропонується нове рішення, що включає адаптивний рівень згортки, який полегшує обмін інформацією між віддаленими словами. Запропонований підхід дозволяє покращити збереження довжини без шкоди для загальної продуктивності.

Напрям виявлення пропаганди на рівні документа висвітлено в роботах [45, 46, 47].

Зокрема, у [45] визнання наявності пропаганди відбувається на двох рівнях: на загальному рівні, тобто на рівні документа, і на рівні окремих речень. Використовуються такі методи побудови ознак, як статистичний індикатор «TF-IDF» [48], модель векторизації «Bag of Words» [49], маркування частин мови, модель «word2vec» [50] для отримання векторних зображень слів, а також розпізнавання тригера-слова (деякі підсилювальні слова, абсолютні займенники та «блискучі» слова). В якості основного алгоритму моделювання використано логістичну регресію, за допомогою якої розпізнавання пропаганди на рівні документів поліпшилося майже в 1,2 раза (на 20%). Аналіз необроблених даних показав, що модель розпізнавання пропаганди на рівні документа змогла

правильно класифікувати 6097 непропагандистських статей і 694 пропагандистські статті. Отримана оцінка моделі: 0,943. Модель розпізнавання пропаганди на рівні речення успішно класифікувала 205 пропагандистських статей і 1917 непропагандистських статей. Оцінка моделі: 0,744 (але 731 статтю було класифіковано неправильно).

У дослідженні [46] запропоновано використовувати точно налаштований, надійно оптимізований підхід попереднього навчання BERT (RoBERTa) і вбудовування слів із використанням класифікації тексту з кількома мітками й кількома класами. Досягнуто точності оцінювання продуктивності 90%, 75%, 68% і 65% відповідно на текстових наборах ProText, PTC, TSHP-17 і Qprop. Метод великих даних, особливо з моделями глибокого навчання [51], може допомогти заповнити незадовільні великі дані в новій стратегії класифікації тексту.

У [47] оцінюється доцільність використання базової технології ChatGPT [52], великих мовних моделей (LLM) [53, 54] для виявлення ознак пропаганди в новинних статтях. Дослідники, розглядаючи напрямки виявлення пропаганди, використовували роботу Мартіно, у якій визначено список із 18-ти різних прийомів пропаганди.

У цьому контексті використання багатовимірних контекстних вбудовувань та гібридних моделей, що поєднують архітектури BiLSTM та трансформери, забезпечує поглиблене розуміння текстового контенту та сприяє підвищенню точності виявлення пропаганди з метою ефективної протидії інформаційним загрозам і забезпечення надійних механізмів виявлення та класифікації пропаганди.

Отже, в межах дисертаційного дослідження для виявлення та класифікації прийомів пропаганди буде використовуватись напрям щодо аналізу контенту на рівні документів. Такий вибір обумовлений тим, що пропаганда, виявлена виключно на рівні фрагментів, не враховує контексту, в той час як

пропагандистський контент характеризується множинністю використовуваних прийомів пропаганди та довгими контекстними семантичними залежностями.

1.3. Аналіз підходів до класифікації прийомів пропаганди

Підхід до виявлення пропаганди визначає, якими інструментами штучного інтелекту аналізуються текстові дані для виявлення прийомів пропаганди [55]. У межах дослідження розглянуто 3 основні підходи щодо класифікації прийомів пропаганди:

– традиційний підхід на основі машинного навчання;
– підхід на основі рекурентних нейромереж;
– підхід на основі моделей-трансформерів.

Традиційний підхід на основі машинного навчання [56, 57] охоплює декілька методів і алгоритмів, призначених для розв'язання різноманітних завдань прогнозування, класифікації та кластеризації даних, зокрема і для виявлення прийомів пропаганди. Лінійна регресія використовується для моделювання лінійних залежностей між вхідними функціями (ознаками) і цільовими значеннями, а також є одним з найпростіших методів регресійного аналізу і часто використовується для прогнозування числових значень. Метод опорних векторів (SVM) шукає оптимальну гіперплощину, яка найкращим чином розділяє два класи точок даних у просторі ознак. SVM часто використовується для задач класифікації, особливо коли дані мають складну структуру [58]. Навчання на основі байєсових мереж ґрунтується на байєсівській ймовірнісній моделі, де кожна змінна розглядається як випадкова, і використовуються правила байєсівського висновку для побудови моделі. Також серед традиційні підходів машинного навчання для виявлення прийомів пропаганди використовується логістична регресія, k-NN, алгоритми кластеризації, наприклад, k-means. У роботі [59] автори наводять результати

дослідження кількох моделей класифікації, включаючи мультиноміальний наївний метод Байєса, SVM [60], логістичну регресію та K-найближчих сусідів. У статті [61] автори представили два альтернативні методи (BERT та SVM) автоматичного визначення прокремлівської пропаганди в газетних статтях і дописах у Telegram.

Підхід до виявлення пропаганди [62] на основі рекурентних нейронних мереж [63] використовується для аналізу послідовних даних, зокрема текстів, що часто зустрічаються у соціальних мережах. Рекурентні нейронні мережі, такі як Long Short-Term Memory і Gated Recurrent Unit [64], – це різновиди архітектур рекурентних нейронних мереж, кожна з яких має свої особливості і застосування у вирішенні різних завдань у машинному навчанні, зокрема у виявленні пропаганди в соціальних мережах. RNN [65] є базовою архітектурою, яка здатна обробляти послідовні дані, зберігаючи інформацію у вигляді внутрішнього стану (пам'яті), який оновлюється при кожному новому вході. LSTM – це розширена версія RNN, яка включає додаткові механізми, такі як ворота забування, ворота оновлення та ворота виходу. GRU [66] є спрощеною версією LSTM [67], яка має менше внутрішніх компонент. GRU вважається менш обчислювально витратною архітектурою порівняно з LSTM [68]. Результати дослідження ідентифікації пропаганди на платформі Twitter під час пандемії COVID-19 авторів [69] свідчать, що запропонована пропагандистська ідентифікація на основі LSTM показала кращі результати, ніж інші розглянуті у роботі методи машинного навчання. За допомогою запропонованого підходу на основі LSTM досягається точність 77.15%. А у статті [70] автори використовують техніки глибокого навчання Bi-LSTM [71] і Bi-GRU [72] з методами SVM із слабким контролем. Даний підхід забезпечив точність 90% у виявленні пропагандиських новин. Автори стверджують, що такий підхід є ефективним і дієвим для нерозмічених даних.

Підхід на основі моделей-трансформерів передбачає використання таких архітектур нейромереж, як BERT, RoBERTa, DistilBERT, GPT тощо [73]. BERT є однією з найвідоміших архітектур трансформерів, розробленою Google. Нейромережа BERT здатна досягати вражаючих результатів у завданнях обробки природної мови (NLP) завдяки здатності до контекстної обробки слів і здібності до підготовки звичайних моделей для багатьох NLP-завдань [74]. RoBERTa – це оптимізований підхід до BERT, який покращує навчання і результати моделі на різних NLP-завданнях шляхом застосування різних оптимізаційних стратегій. DistilBERT [75] вважається легковаговою версією BERT, яка має сутність оригінальної моделі, зменшуючи кількість параметрів і зберігає високу продуктивність на різних NLP-завданнях. GPT [76] – це родина моделей трансформерів, розроблених OpenAI [77]. Даний підхід застосовують у дослідженні для класифікації пропаганди [78]. Автори використовують три моделі глибокого навчання (CNN, LSTM, Bi-LSTM) і чотири багатомовні моделі на основі трансформерів, а саме: BERT, Distil-BERT, Hindi-BERT і Hindi-TPU-Electra. Експериментальні результати вказують на те, що багатомовні моделі BERT і Hindi-BERT забезпечують найкращу продуктивність із найвищим показником F1 84% за даними проведеного експерементального дослідження. Також у роботі [79] досліджується продуктивність BERT [80] і RoBERTa [81], DeBERTa [82] з комбінацією різних методів збільшення даних для виявлення пропагандистських текстів. Автори змогли досягти F1 micro 60% на тестовому наборі, використовуючи ансамбль моделей BERT, RoBERTa та DeBERTa.

Результати проведеного аналізу свідчать, що всі наведені підходи знаходять своє застосування у вирішенні задач виявлення прийомів пропаганди. Оскільки кожний з підходів має свої особливості застосування, то в межах дослідження буде порівняна їх ефективність для класифікації прийомів пропаганди безпосередньо в контексті запропонованих рішень.

1.4. Проблеми виявлення та класифікації прийомів і об’єктів пропаганди

Пропаганда у текстових даних виявляється шляхом визначення множини семантичних ознак (маркери, прийоми, об’єкти), які перебувають у взаємозв’язку та когнітивно формують складну систему, для оцінювання якої вимагається аналіз як кожної з ознак, так і існуючих залежностей між ними. Це визначає ряд проблем, що ускладнюють вирішення задачі виявлення та класифікації прийомів і об’єктів пропаганди.

Одною з головних проблем щодо виявлення пропаганди є потреба в наявності достовірно маркованих даних. Щоб допомогти науковому співтовариству ідентифікувати пропаганду в текстових новинах, у дослідженні [83] запропоновано бібліотеку пропагандистських текстів (ProText). Мітки правдивості призначаються сховищам ProText [84] після ручної та автоматичної перевірки за допомогою методів перевірки фактів. Авторами було використано підхід до обробки природної мови, щоб створити систему, яка використовує глибоке навчання для автоматичного визначення пропаганди в новинах. У [85] представлено багатомовний корпус із 12 000 дописів у Facebook, повністю анотованих щодо упередженості та пропаганди. Корпус був створений як частина спільного завдання «FigNews 2024» щодо наративів у ЗМІ для створення кадрів ізраїльської війни проти Гази. Він охоплює різні події під час війни з 7 жовтня 2023 року по 31 січня 2024 року. Корпус містить дописи п’ятьма мовами: арабською, івритом, англійською, французькою та гінді, по 2 400 дописів для кожної мови. У процесі анотування брали участь 10 аспірантів спеціальності «Правознавство». Угода між анотаторами (IAA) використовувалася для оцінки анотацій корпусу із середнім показником IAA [86, 87] 80.8% для упередженості та 70.15% для пропагандистських анотацій.

За результатом аналізу пов’язаних робіт у сфері виявлення прийомів та об’єктів пропаганди [88] виявлено такі проблеми: відсутність комплексного аналізу взаємозв’язків прийомів та об’єктів пропаганди в текстах; відсутність узагальнень для об’єктів пропаганди та їх альтернативних згадувань в текстах. Пропаганда, детектована шляхом пошуку лише іменованих сутностей, не має наочності щодо спрямованості пропаганди із застосуванням прийомів; однак прийоми пропаганди, детектовані на рівні документа, не відображають об’єктів спрямування пропаганди. Водночас при виявленні пропаганди як завдання виявлення іменованих сутностей (NER) [89] виникає проблема: об’єкти пропаганди подаються не власними назвами і навіть не синонімами, а їх семантична близькість не відображається очевидним чином.

Ще однією проблемою у виявленні і класифікації прийомів та об’єктів пропаганди у текстовому контенті засобами ШІ є відсутність інтерпретації рішень, отриманих засобами ШІ. Для усунення ефекту «чорного ящика» існують методи візуальної аналітики [31], серед яких є LIME [43].

LIME є методом пояснення прогнозів моделей машинного навчання, який фокусується на локальній інтерпретації рішень. Його основна ідея полягає в апроксимації складної моделі простою лінійною моделлю в околі конкретного передбачення. Для цього LIME створює штучний локальний набір даних, що включає варіації вихідних характеристик та обчислює відповіді моделі. Далі застосовується вагова регресія, яка враховує відстань змінених зразків до вихідної точки, що дозволяє ідентифікувати внесок окремих ознак у конкретне передбачення. Це забезпечує зрозумілу та локальну інтерпретацію навіть для складних алгоритмів, зберігаючи при цьому незалежність від конкретного типу моделі. Метод широко використовується для пояснення рішень глибоких нейронних мереж, ансамблевих методів та інших нелінійних моделей. Приклад застосування пояснень інтерпретаційною моделлю LIME наведено на рис. 1.1.

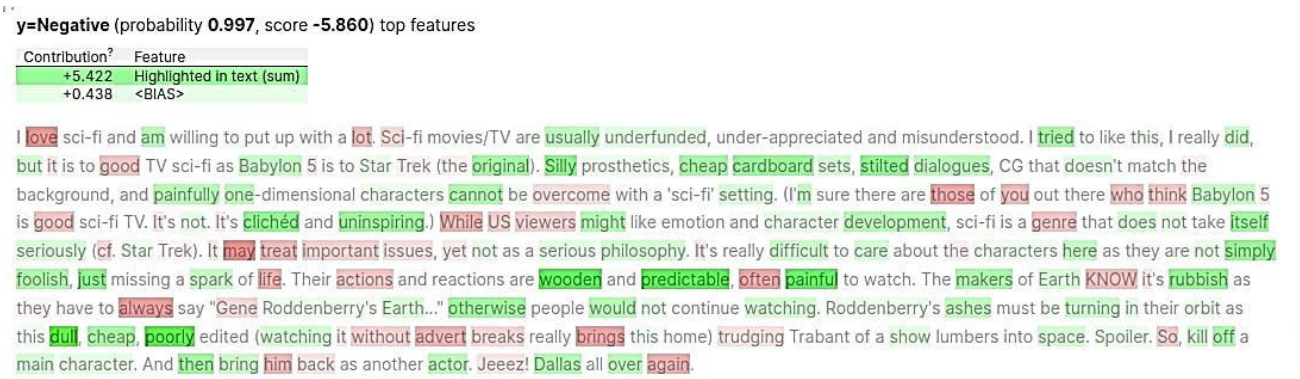


Рис. 1.1 – Приклад пояснення рішення інтерпретаційною моделлю LIME [90]

Відповідно до рис. 1.1 зелену підсвітку мають слова, які впливають на позитивне рішення віднесення зразку до цільового класу, а червоні – до нецільового.

Отже, на поточний момент існуючи методи та засоби мають ряд невирішених проблем, зокрема, низьку точність класифікації прийомів пропаганди, відсутність поясненості існуючих рішень. Також окремими проблемами є відсутність комплексного аналізу взаємозв'язків прийомів та об'єктів пропаганди в текстах, відсутність узагальнень для об'єктів пропаганди та їх альтернативних згадувань у текстах. Відповідно дисертаційне дослідження має на меті вирішення наведених проблем.

1.5. Визначення термінології предметної області пропаганди та правове регулювання пропаганди в Україні

Визначення термінології предметної області пропаганди та аналіз правового регулювання пропаганди дозволяє як встановити значущість проблеми виявлення та класифікації пропаганди на державному рівні, так і забезпечити формування об'єктивного подання сутності пропаганди в контексті національних інтересів для коректного підбору методів та засобів для виявлення та класифікації пропаганди в межах дослідження.

Пропаганда визначається як систематичне поширення ідей, поглядів або наративів, спрямованих на формування суспільної думки, політичних переконань або поведінкових моделей [91]. Вона може використовуватися як інструмент інформаційного впливу з боку державних чи недержавних «акторів».

Сучасна наукова література трактує пропаганду як цілеспрямований інформаційний вплив, який може здійснюватися політичними, державними або іншими організаціями.

Законодавче визначення пропаганди в Україні є фрагментарним і залежить від конкретного контексту її застосування. В українському законодавстві окремо визначено заборону на пропаганду фашизму, нацизму (Проєкт Закону України від 12.11.2009 № 5247-1 [91]) та російського нацистського тоталітарного режиму (Закон України від 22.05.2022 № 2265-IX [92]), що безпосередньо пов’язано з питаннями національної безпеки та інформаційного захисту.

Закон України «Про заборону пропаганди російського нацистського тоталітарного режиму, збройної агресії Російської Федерації як держави-терориста проти України, символіки воєнного вторгнення російського нацистського тоталітарного режиму в Україну» встановлює чіткі нормативно-правові засади щодо обмеження пропаганди, зокрема в письмових формах, таких як публікації, статті та інші текстові матеріали. Визначення пропаганди, згідно з першим пунктом статті 1 цього Закону, охоплює поширення інформації, що підтримує або виправдовує злочинну діяльність Російської Федерації та її органів, популяризацію діяльності органів держави-терориста та їх представників, а також публічне заперечення злочинного характеру збройної агресії Російської Федерації проти України, включаючи поширення такої інформації через медіа та інтернет. Також пропагандою є використання символіки воєнного вторгнення російського нацистського тоталітарного

режиму, а також популяризація ідеології «руского міра» у будь-який спосіб і засобами.

Закон встановлює пряму заборону на пропаганду, що включає в себе заборону на використання зазначеної символіки воєнного вторгнення, зокрема таких символів, як латинські літери «Z» та «V», якщо вони застосовуються без правомірного контексту або у контексті виправдання збройної агресії, а також на використання символіки збройних сил Російської Федерації. Заборона поширюється на публічний простір, включаючи друковані матеріали, рекламу в медіа, інтернеті та соціальних мережах. Водночас закон передбачає певні винятки: правомірне використання символіки у творах, спрямованих на засудження російського нацистського тоталітарного режиму та його агресії, а також у наукових дослідженнях та освітніх матеріалах, які використовуються в освітньому процесі.

Пропаганда визначається як текстова інформація, що містить маніпулятивні прийоми, спрямовані на виправдання або популяризацію ідеологій, заборонених законодавством України.

Отже, законодавство України чітко розмежовує правомірну інформаційну діяльність і заборонену пропаганду, зокрема в контексті гібридних загроз та інформаційних війн. Таким чином, розуміння сутності пропаганди та її правового регулювання є ключовим для аналізу механізмів впливу інформаційних кампаній на суспільство. У межах дослідження термін «пропаганда» розглядається у контексті політичної та військової тематики, що включає інформаційні кампанії, спрямовані на формування громадської думки щодо державної політики, міжнародних конфліктів, воєнних дій або безпекових аспектів. Доцільним є подання пропаганди як системи семантичних ознак тексту, оцінювання кожної з яких та їх взаємозалежностей забезпечить вирішення задачі, виявлення та класифікації пропаганди в текстових джерелах.

1.6. Класифікація пропаганди

Виявлення та класифікація прийомів і об’єктів пропаганди у текстових даних потребує виявлення множини семантичних ознак (види, маркери, прийоми, об’єкти, взаємозв’язки тощо). Ці семантичні ознаки описово відомі та мають бути поданими у вигляді системи класифікацій для забезпечення можливості їх автоматизованого оцінювання в межах вирішення задач дослідження.

Наукова спільнота з обробки природної мови виявляє зростаючий інтерес до виявлення конкретних прийомів пропаганди, а також визначення конкретних ділянок кожного випадку. Цей інтерес виявляється через організацію спільних завдань, таких як NLP4IF-2019 на тему тонкої детекції пропаганди [93], завдання SemEval-2020 11 на виявлення прийомів переконання в новинних статтях [28, 94], завдання SemEval-2021 6 на виявлення прийомів переконання в текстах і зображеннях [95], завдання WANLP-2022 на виявлення пропаганди арабською мовою [96] і завдання SemEval-2023 Task 3: виявлення категорій, фреймів і прийомів переконання в онлайн-новинах у багатомовному середовищі [97]. Завдання 3 – «Прийоми переконання» в межах CheckThat! Lab 2024 [98] є черговою спробою розвинути сучасні дослідження в цьому напрямі.

Пропаганда може бути класифікована за типами залежно від джерела та ступеня відкритості щодо її походження. Існує три основні типи пропаганди: біла, сіра і чорна [99].

Під білою пропагандою [100] вважають найбільш прозорий тип пропаганди, який чітко вказує на джерело інформації та його наміри. Білу пропаганду часто використовують уряди та офіційні організації для поширення позитивних повідомлень, які підкреслюють переваги певної політики або дій. Як приклад, використання плакатів під час воєн, що закликають до патріотизму та підтримки уряду.

Сіра пропаганда [101] відрізняється тим, що її джерело або не є чітко визначеним, або ж залишається прихованим. Цей тип пропаганди може використовуватися для поширення напівправди або змішаної інформації, яка не має явного джерела, що ускладнює перевірку її достовірності. Наприклад, повідомлення, які створюють амбівалентність щодо походження інформації, часто використовуються в політичних кампаніях для підриву довіри до опонентів.

Чорна пропаганда [102] є найбільш оманливою формою, оскільки видає себе за повідомлення, що походять від одного джерела, тоді як насправді воно створене іншим, часто ворожим джерелом. Цей тип пропаганди спрямований на дискредитацію, поширення дезінформації [103, 104] та створення негативного іміджу об’єкта спрямування. Наприклад, підробка документів або фальшивих заяв, які нібито походять від противника, але насправді створені для введення в оману.

Класифікація пропаганди на білу, сіру та чорну [105] базується на рівні прозорості джерела та правдивості інформації. Ця традиційна класифікація допомагає зрозуміти, як різні прийоми пропаганди використовуються для досягнення певних цілей. Перехід від цієї трирівневої класифікації до більш детального розгляду конкретних прийомів пропаганди обґрунтований необхідністю аналізу та ідентифікації маніпулятивних прийомів пропаганди [106, 107], що використовуються в сучасному інформаційному середовищі. Оскільки біла пропаганда відкрито ідентифікує своє джерело та має чітку мету, часто спрямовану на позитивний вплив, у роботі розглядатись на досліджуватись не буде, а основна увага буде приділена виявленню та класифікації сірої та чорної пропаганди.

Сіра пропаганда, яка частково приховує або взагалі не розкриває своє джерело, створює невизначеність та плутанину. Вона часто застосовує прийоми, що викликають сумніви або перебільшують факти для маніпуляції сприйняттям

аудиторії. Ця невизначеність дозволяє сірій пропаганді бути більш гнучкою у досягненні своїх цілей, використовуючи різні риторичні прийоми без чіткої вказівки на своє походження.

Чорна пропаганда є найбільш маніпулятивною та оманливою, оскільки її джерело зазвичай маскується або навмисно вводить в оману та має спрямованість на дискредитацію та дезінформацію.

Розширена класифікація конкретних прийомів пропаганди дозволяє більш детально аналізувати механізми впливу та маніпуляції. У [108] запропоновано поділ за категоріями прийомів пропаганди (рис. 1.2).

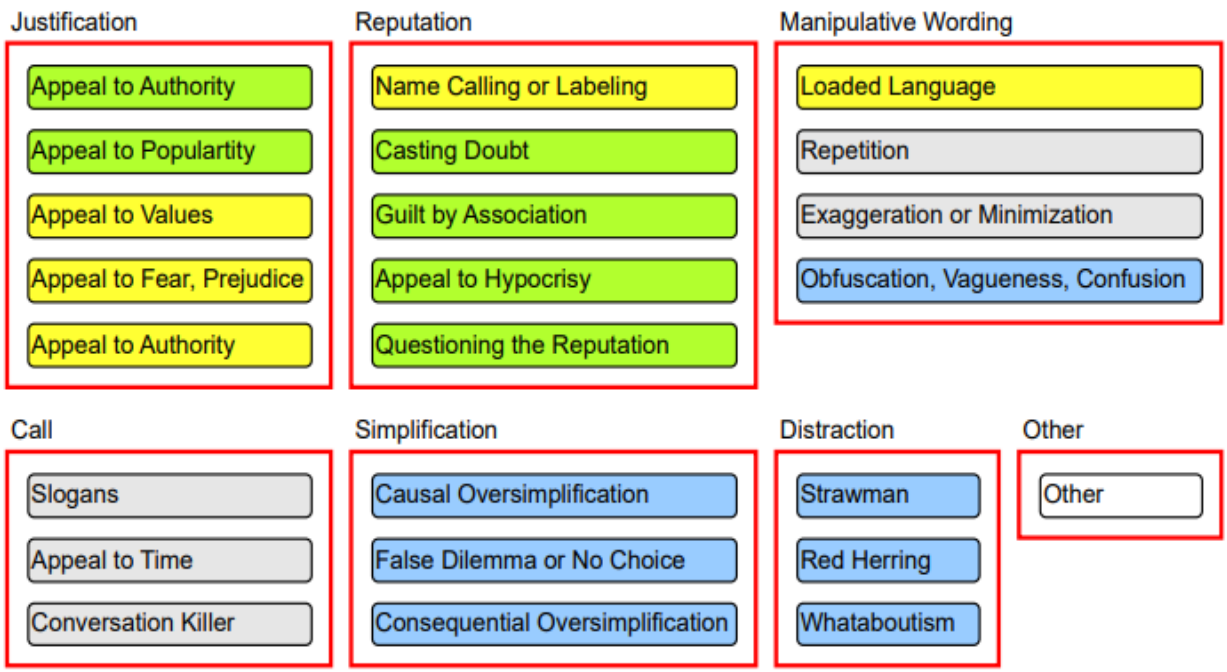


Рис. 1.2 – Таксономія прийомів пропаганди за [108]

Прийоми пропаганди у текстовому контенті характеризуються певними семантичними маркерами, за якими можна оцінити приналежність тексту до того чи іншого прийому пропаганди. Нижче наведено систему класифікації прийомів пропаганди та їх опис, запропоновану у [41]. Далі в роботі під класифікацією прийомів пропаганди матиметься на увазі описані нижче

прийоми, а під маркерами - «емоційність тексту», «булінг», «страх» та «мова ворожнечі».

«Апеляція до авторитету» [106] – прийом, при якому аргумент підкріплюється посиланням на думку або твердження авторитетної особи чи організації без критичного аналізу. Наприклад, «Цей препарат рекомендує відомий лікар, отже він ефективний».

«Помилка хибної дихотомії» [107] – подання ситуації так, наче існують лише два протилежних варіанти, без врахування інших можливих. Наприклад, «Або ви підтримуєте нас, або ви наш ворог».

«Сумнів» [109] – це формування недовіри або підозри без надання конкретних доказів. Наприклад, «Ви дійсно вірите, що вони говорять правду?».

«Причинно-надмірна спрощеність» [110] – це пояснення складних явищ або подій надто простими причинами. Наприклад, «Безробіття зросло, тому що люди стали лінивими».

«Перебільшення» [111] – це надання подіям, фактам або явищам більшого значення, ніж вони мають насправді. Наприклад, «Цей проєкт змінить життя мільйонів людей назавжди».

«Розмахування прапором» [112] – це звернення до патріотичних почуттів, щоб підтримати аргумент або викликати емоційну реакцію. Наприклад, «Це рішення найкраще для нашої країни, справжні патріоти підтримають його».

«Заряджена мова» [113] – це використання слів і фраз з сильним емоційним забарвленням для впливу на аудиторію. Наприклад, «зрада», «героїзм», «катастрофа».

«Навішування ярликів» – це застосування негативних або позитивних стереотипів для дискредитації або підтримки певної особи, групи або ідеї. Наприклад, «Цей політик – справжній демагог».

«Мінімізація» – це зменшення значущості або серйозності події чи проблеми. Наприклад, «Це всього лише дрібна неприємність, не варто звертати увагу».

«Обзивання» – це використання образливих або принизливих слів для дискредитації опонента. Наприклад, «Ці люди – невігласи».

«Зведення до Гітлера» – це порівняння опонента або його аргументів з Гітлером чи нацистами з метою дискредитації. Наприклад, «Це як у Гітлера, значить, це неправильно».

«Червоний оселедець» – це введення в аргументацію несуттєвого факту або питання, щоб відволікти від основної теми. Наприклад, «Ми говоримо про економіку, але давайте згадаємо, як він поводився в особистому житті».

«Повторення» – це багаторазове повторення одного і того ж аргументу або твердження для закріплення його в свідомості аудиторії. Наприклад, «Ми повторюємо: це рішення найкраще для всіх».

«Гасла» [113] – це короткі, вражаючі фрази, що легко запам'ятовуються і мають сильний емоційний вплив. Наприклад, «За мир і справедливість!».

«Ватабоутізм» [114] – це відволікання від критики шляхом вказування на інші проблеми або помилки, часто не пов'язані з основною темою. Наприклад, «Ви критикуєте нашу політику, а як щодо вашої країни?».

«Кліше, що припиняють думання» – це використання загальновідомих фраз або кліше, що закінчують дискусію або критику. Наприклад, «Такі справи, життя несправедливе».

«Апеляція до страху» – це прийом, при якому аргумент підкріплюється викликанням страху або упереджень у аудиторії з метою маніпуляції її думками чи діями. Наприклад, «Якщо ми не посилимо заходи безпеки, наша країна може стати наступною мішенню терористів». Такий підхід може переконати людей

діяти певним чином через емоційний вплив, а не через логічні аргументи або фактичні докази.

Такий вибір маркерів ґрунтується на їхній функціональній ролі у пропагандистських текстах та їхньому впливі на когнітивні й емоційні процеси аудиторії. Емоційність тексту є ключовим інструментом маніпуляції, оскільки емоції мають сильніший вплив на сприйняття, ніж раціональний аналіз. Висока емоційна напруженість повідомлення знижує критичне мислення, підвищує когнітивне навантаження та сприяє прийняттю інформації без її перевірки. Булінг як комунікативна стратегія слугує засобом соціального тиску. Він формує негативне ставлення до певних груп чи осіб, сприяючи конформізму та соціальному виключенню опонентів. Це механізм, що використовується для обмеження плюралізму думок через страх негативної оцінки. Страх є одним із найефективніших механізмів впливу на поведінку, оскільки активізує захисні реакції та знижує здатність до раціонального аналізу. Використання страху у повідомленнях спрямоване на викликання тривоги, що, у свою чергу, підвищує готовність до прийняття радикальних рішень або підтримки авторитарних заходів. Мова ворожнечі є засобом дегуманізації опонента, що полегшує виправдання агресії, дискримінації чи соціальної ізоляції. Вона знижує рівень емпатії та нормалізує насильницькі наративи, що сприяє створенню атмосфери конфлікту та поляризації.

Таким чином, ці маркери не лише систематизують пропагандистські прийоми, а й пояснюють їхній психологічний та соціальний вплив, що робить їх релевантними для дослідження маніпулятивних технологій.

Відповідно до проведеного аналізу відомих методів виявлення пропаганди, результати можна навести у таблиці 1.1.

Таблиця 1.1

Порівняння існуючих підходів до виявлення пропаганди за метриками

Назва методу, автори:	**Accuracy, %**	**Precision, %**	**Recall, %**	**F1, %**
Логістична регресія (*Висоцька*) [45]	94.33	не вказано	не вказано	не вказано
GPT-4 base *(Kilian Sprenkamp, Daniel Gordon, Liudmila Zavolokina)* [94]	не вказано	52.86	64.52	58.11
GPT-4 chain of thought *(Kilian Sprenkamp, Daniel Gordon, Liudmila Zavolokina)* [94]	не вказано	56.86	57.82	57.34
XLM RoBERTa Large зі стилометрією (*Aleš Horák, Radoslav Sabol, Ondřej Herman, Vít Baisa*) [35]	не вказано	не вказано	не вказано	92.26
XLM RoBERTa Large без стилометрії (*Aleš Horák, Radoslav Sabol, Ondřej Herman, Vít Baisa*) [35]	не вказано	не вказано	не вказано	91.07

Відповідно до наведеного опису відомих прийомів пропаганди, факт використання кожного з них визначається за проявом множини характерних для цього прийому маркерів. Це відкриває можливість для їх автоматизованого виявлення засобами ШІ.

1.7. Мета та завдання дослідження

На основі виконаного аналізу було помічено, що хоч сучасні засоби ШІ є досить ефективними для автоматизованого виявлення та класифікації прийомів і об’єктів пропаганди, однак, попри бурхливий розвиток галузі обробки природної мови, відсутній комплексний інструмент, який би не лише забезпечував виявлення та класифікацію прийомів пропаганди, а й показував на кого і на що вона спрямована. Також проблемою використання моделей

глибокого навчання є їх низька інтерпретованість, що може знижувати довіру до результатів нейромережевого виявлення пропаганди.

Метою роботи є підвищення точності та якості виявлення прийомів та об’єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень. Основна ідея полягає у створенні методу на основі моделі гібридної архітектури для виявлення пропаганди; створенні методу виявлення прийомів пропаганди за семантичними маркерами на основі набору моделей глибокого навчання, навчених на модифікованих маркованих даних окремо для кожного прийому пропаганди; створенні методу для вирішення проблеми ідентифікації об’єктів пропаганди, який дозволяє знаходити в пропагандистських текстах на кого і на що спрямовані пропагандистські прийоми.

Для досягнення мети роботи необхідним є виконання таких *завдань*:

1. Провести аналіз методів, засобів та технологій для автоматизованого виявлення пропаганди у текстовому контенті.
2. Удосконалити метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання.
3. Розробити метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень.
4. Розробити метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.
5. Розробити інтелектуальну інформаційну систему для валідації запропонованих методів і провести експериментальні дослідження.

1.8. Висновки до розділу 1

У першому розділі дисертації проведено комплексний аналіз методів, засобів та технологій для автоматизованого виявлення пропаганди у текстовому

контенті. Актуальність застосування засобів штучного інтелекту для виявлення та класифікації прийомів пропаганди підтверджується необхідністю ефективного протистояння сучасним інформаційним загрозам. Проведений аналіз підтверджує важливість подальшого розвитку та вдосконалення методів штучного інтелекту для автоматизованого виявлення пропаганди, що є критичним для забезпечення інформаційної безпеки та протидії маніпулятивному впливу на суспільство.

Проаналізовано існуючі підходи до виявлення та класифікації прийомів пропаганди, що дозволило виокремити семантичні особливості, притаманні прийомам. Проаналізовано наявні системи класифікації пропагандистських прийомів, що демонструють різний рівень точності, та на сучасному етапі показник точності потребує покращення.

Виявлено низку проблем, пов'язаних з виявленням та класифікацією пропагандистських прийомів, зокрема, складність автоматизації процесу через різноманіття мовних конструкцій та контекстів. Запропонована класифікація прийомів пропаганди на основі сучасних досліджень у цій галузі. У межах дисертаційного дослідження буде виконуватись класифікація за 17-ма прийомами пропаганди: «Апеляція до авторитету», «Помилка хибної дихотомії», «Сумнів», «Причинно-надмірна спрощеність», «Перебільшення», «Розмахування прапором», «Заряджена мова», «Навішування ярликів», «Мінімізація», «Обзивання», «Зведення до Гітлера», «Червоний оселедець», «Повторення», «Гасла», «Ватабоутізм», «Кліше, що припиняють думання», «Апеляція до страху».

За результатом проведеного аналізу методів, засобів та технологій для автоматизованого виявлення пропаганди метою роботи визначено підвищення точності та якості виявлення прийомів та об’єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень.

РОЗДІЛ 2.
МЕТОД ВИЯВЛЕННЯ ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ

2.1. Концепція виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті

Концепція роботи щодо виявлення і класифікації прийомів та об’єктів пропаганди у текстовому контенті з використанням нейромережевих засобів передбачає три послідовних етапи та забезпечує автоматизоване виявлення наявності пропаганди, класифікацію використаних прийомів пропаганди, встановлення об’єктів виявленого пропагандистського впливу та візуальну аналітику одержаних результатів [42, 115], як наведено на рис. 2.1.

Рис. 2.1 – Концепція нейромережевого виявлення і класифікації прийомів та об’єктів пропаганди

Вхідними даними запропонованої концепції є текст для аналізу, навчена модель глибокого навчання для бінарної класифікації (пропаганда / не пропаганда), множина нейромережевих моделей глибокого навчання для ідентифікації прийомів пропаганди (окрема нейромережа для кожного з 17-ти прийомів пропаганди).

Метод 1 (наведено в п. 2.5) призначений для класифікації текстів за вмістом пропаганди. Метод здійснює аналіз рівня пропаганди та відносить її до

однієї з категорій: «без пропаганди», «підозрілий», «пропагандистський». Віднесення до категорій здійснюється відповідно до отриманої нейромережевої оцінки рівня пропаганди.

Метод 2 (наведено в п. 3.2) відповідає за виявлення прийомів пропаганди та їх візуальну інтерпретацію. Застосовується лише для текстів, що віднесені в категорію «пропагандистський текст». Вхідне текстове представлення подається по черзі на 17 навчених нейромережевих моделей для аналізу наявності 17-ти прийомів. Відповідно вихідними даними буде нейромережева оцінка наявності прийомів пропаганди за маркерами [43, 106] та подання з візуальною інтерпретацією результатів [107].

Метод 3 (наведено в п. 3.3) відповідає за виявлення об’єктів пропаганди та їх візуальну інтерпретацію. Використовує результати, отримані в ході виконання попередніх методів. Здійснює перетворення вхідних даних у множину тематичних об’єктів пропаганди із взаємозв’язками виявлених об’єктів з прийомами пропаганди.

На прикладі типового тексту наведено застосування представленої концепції (текст надано експертом): «*Останні події чітко доводять, що наші вороги діють злагоджено та рішуче. Вони не зупиняться, поки не знищать усе, що ми створили за роки незалежності. Західні союзники вже неодноразово показували свою слабкість і байдужість до нашої долі, тож розраховувати можемо тільки на себе. Поки ми сумніваємось – нас нищать. Поки ми чекаємо – ворог укріплюється. Єдиний шлях – це мобілізація всіх ресурсів і негайна дія. Кожен, хто у цей час мовчить або критикує, – грає на руку ворогу. Ми повинні очистити наші ряди від зрадників і слабких, бо інакше катастрофа неминуча. Якщо ми зараз не підніземось на захист нашої Батьківщини, завтра буде пізно. Ворог не чекає, і ми не маємо права зволікати. Годі терпіти! Вставай, народе! Настав час діяти – радикально і безжально!*».

Метод 1, який відповідає за класифікацію текстів за вмістом пропаганди, дасть нейромережеву оцінку 0.85, що характеризує текст як пропагандистський.

Оскільки текст класифіковано як пропагандистський, то далі цей текст аналізується в межах методу 2, який відповідає за виявлення прийомів пропаганди та візуальну інтерпретацію. Згідно з текстом було отримано такі оцінки за прийомами пропаганди: «Апеляція до авторитету» – виражено на 0.10, «Помилка хибної дихотомії» – виражено на 0.81, «Сумнів» – виражено на 0.40, «Причинно-надмірна спрощеність» – виражено на 0.45, «Перебільшення» – виражено на 0.92, «Розмахування прапором» – виражено на 0.87, «Заряджена мова» – виражено на 0.73, «Навішування ярликів» – виражено на 0.85, «Мінімізація» – виражено на 0.05, «Обзивання» – виражено на 0.78, «Зведення до Гітлера» – виражено на 0.02, «Червоний оселедець» – виражено на 0.18, «Повторення» – виражено на 0.36, «Гасла» – виражено на 0.60, «Ватабоутізм» – виражено на 0.15, «Кліше, що припиняють думання» – виражено на 0.10, «Апеляція до страху» – виражено на 0.87. Відповідно за порогом 0.5 виявленими прийомами вважаються: «Помилка хибної дихотомії (0.81), «Перебільшення» (0.92), «Розмахування прапором» (0.87), «Заряджена мова» (0.73), «Навішування ярликів» (0.85), «Обзивання» (0.78), «Гасла» (0.60), «Апеляція до страху» (0.87).

Приклад візуальної інтерпретації для прийому «Апеляція до страху» (0.87): «*Останні події чітко доводять, що наші вороги діють злагоджено та рішуче. Вони не зупиняться, поки не знищать усе, що ми створили за роки незалежності. Західні союзники вже неодноразово показували свою слабкість і байдужість до нашої долі, тож розраховувати можемо тільки на себе. Поки ми сумніваємось – нас нищать. Поки ми чекаємо – ворог укріплюється. Єдиний шлях – це мобілізація всіх ресурсів і негайна дія. Кожен, хто у цей час мовчить або критикує, – грає на руку ворогу. Ми повинні очистити наші ряди від зрадників і слабких, бо інакше катастрофа неминуча. Якщо ми зараз не*

підніземось на захист нашої Батьківщини, завтра буде пізно. Ворог не чекає, і ми не маємо права зволікати. Годі терпіти! Вставай, народе! Настав час діяти – радикально і безжально!».

Метод 3 відповідає за виявлення об'єктів пропаганди та їх візуальну інтерпретацію. У тексті виявлені такі об'єкти пропаганди: «*ворог*», «*вони*», «*ми*», «*Західні союзники*», «*нас*», «*зрадники*», «*захист*», «*Батьківщина*», «*народе*». Знайдені об'єкти та близькі їм семантичні подання наведено нижче:

«*Останні події чітко доводять, що наші вороги діють злагоджено та рішуче. Вони не зупиняться, поки не знищать усе, що ми створили за роки незалежності. Західні союзники вже неодноразово показували свою слабкість і байдужість до нашої долі, тож розраховувати можемо тільки на себе. Поки ми сумніваємось – нас нищать. Поки ми чекаємо – ворог укріплюється. Єдиний шлях – це мобілізація всіх ресурсів і негайна дія. Кожен, хто у цей час мовчить або критикує, – грає на руку ворогу. Ми повинні очистити наші ряди від зрадників і слабких, бо інакше катастрофа неминуча. Якщо ми зараз не підніземось на захист нашої Батьківщини, завтра буде пізно. Ворог не чекає, і ми не маємо права зволікати. Годі терпіти! Вставай, народе! Настав час діяти – радикально і безжально!*».

Наведений підхід до нейромережевого виявлення і класифікації прийомів та об'єктів пропаганди є комплексним та дозволяє не лише ідентифікувати наявність пропаганди, а й забезпечити візуальну інтерпретацію отриманих результатів.

2.2. Етичні аспекти та відповідальність впровадження концепції виявлення та класифікації прийомів та об'єктів пропаганди

Пояснювальний ШІ має складну й суперечливу природу та виник як відповідь на потребу підвищення прозорості в алгоритмічних рішеннях [116].

Ідея пояснювального ШІ полягає в тому, що зрозумілість роботи алгоритмів має сприяти зростанню довіри, забезпеченню нагляду та зменшенню ризиків.

Однак попри благородну мету, прагнення до пояснюваності супроводжується глибшими викликами, особливо в контексті підзвітності. У простішому розумінні прозорість означає спробу демістифікувати внутрішні механізми прийняття рішень у ШІ [42], однак ця спрощеність часто створює лише ілюзію розуміння. Застосування таких інструментів, як LIME чи SHAP, дозволяє інтерпретувати складні моделі, але водночас знижує точність і може призводити до хибних тлумачень. З цього виходить, що прозорість не гарантує підзвітності, адже остання передбачає не лише знання про те, як ухвалюється рішення, а й чому воно ухвалюється, з урахуванням упереджень і мотивацій, що формують результат.

Ще однією проблемою є етичні виклики, пов’язані з обмеженнями прозорості. Алгоритми, які формально є пояснюваними, можуть приховувати дискримінаційні наслідки, якщо в них не враховано інклюзивність та етичні аспекти при розробці. Є випадки, коли системи розпізнавання облич демонструють вищу похибку щодо людей із темнішим кольором шкіри та жінок, що свідчить про структурні упередження.

Тому прозорість має бути не кінцевою метою, а лише відправною точкою для досягнення справжньої підзвітності. Це потребує регуляторних рамок, етичного контролю на всіх етапах розробки ШІ та міждисциплінарної співпраці. Лише тоді можливо створити справді відповідальні технології, які слугуватимуть суспільному добру.

У межах концепції виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті необхідно зазначити, що дана задача пов’язана з політичними, соціальними чи культурними контекстами, де ризик неправомірного маркування або викривлення інформації особливо високий.

Небезпека прийняття рішень виключно ШІ полягає в тому, що системи ШІ можуть успадковувати або навіть посилювати людські упередження, закладені в навчальні дані. Це створює ризики для свободи слова, маніпуляції суспільною думкою або дискримінації певних груп. Тому під час розробки та впровадження таких систем важливо не лише забезпечувати високу точність класифікації, а й впроваджувати прозорі механізми пояснюваності рішень штучного інтелекту людиною. Дане дослідження є допоміжною ланкою у виявленні та класифікації прийомів і об’єктів пропаганди, вся відповідальність покладається на людину, яка буде працювати з розробленими методами та відповідними їм програмними реалізаціями.

Відповідно до етичних аспектів запропонованих методів, вони мають високий рівень відтворюваності (досягається детальними описами методів та джерел даних) та прозорості (досягається візуальними поданнями та візуальною поясненістю прийнятих нейромережами рішень).

2.3. Модель семантичної структури пропаганди

Відповідно до запропонованої в п.2.1 концепції виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті, яка використовує нейромережеві засоби та дозволяє виявляти приналежність об’єктів до використаних прийомів пропаганди, наведено формалізоване подання [123] семантичної моделі пропаганди *SMP* (2.1) у тестовому тексті *T*:

$$SMP = CP \cup VP \cup MM \cup TT' \cup PO' \cup RTO' \cup LA \cup Metadata, \quad (2.1)$$

де *CP* – клас тексту *T* за рівнем пропаганди («без пропаганди», «пропагандистський» або «підозрілий»); *VP* – загальна відсоткова оцінка наявності пропаганди; *MM* – множина семантичних маркерів пропаганди з визначеними оцінками прояву кожного з них; *TT′* – множина використаних прийомів ($TT' \subset TT$, де *TT* – множина усіх можливих прийомів пропаганди) з оцінкою їх прояву; *PO′* – множина слів, що представляють пропагандистські

об'єкти в тестовому тексті T з їх семантичною оцінкою ($PO' \subset PO$, де PO – загальна множина об’єктів, виявлених за NER); RTO' – множина важливих зв’язків між прийомами TT' і об’єктами PO' з оцінкою їх семантичної важливості; LA – множина візуальних подань-пояснень, виконаних відповідними моделями-інтерпретаторами; *Metadata* – комплекс допоміжних компонентів.

До комплексу допоміжних компонентів *Metadata* у (2.1) належать складові, які не є інформативним відображенням моделі пропаганди у тексті, але використовуються для одержання вихідних даних (2.2).

$$Metadata = T' \cup DM \cup FT \cup NM' \cup CW \cup CW' \cup LI, \quad (2.2)$$

де T' – попередньо оброблений тестовий текст [124, 125]; DM – навчена нейромережева модель глибокого навчання для оцінки наявності пропаганди у тексті (VP); FT – множина попередньо навчених нейромережевих моделей для виявлення кожного з маркерів пропаганди; NM' – множина попередньо навчених нейромережевих моделей для виявлення кожного прийому пропаганди [129], отриманих шляхом тонкої настройки відповідних моделей з множини NM; CW – множина контекстних вікон, які містять кожну появу об’єктів представлення слів у тестовому тексті T; CW' – множина об’єднаних контекстних вікон, що асоційована з кожним словом представлення об'єктів у тестовому тексті T для нейромережевого аналізу; LI – множина моделей-інтерпретаторів для створення візуальних пояснень LA відповідних моделей множини NM'.

Як вже було вище зазначено, у дисертаційному дослідженні використано класифікацію за 17-ма прийомами пропаганди, відповідно множина TT прийме вигляд (2.3):

$$TT = \{\text{«Апеляція до авторитету», «Помилка хибної дихотомії», «Сумнів», «Причинно-надмірна спрощеність», «Перебільшення», «Розмахування прапором», «Заряджена мова», «Навішування ярликів», «Мінімізація», «Обзивання», «Зведення до Гітлера», «Червоний оселедець», «Повторення», «Гасла», «Ватабоутізм», «Кліше, що припиняють думання», «Апеляція до страху»}.} \quad (2.3)$$

Під додатковою множиною маркерів слід розуміти використання різноманітних ознак в тексті, які притаманні визначеним прийомам пропаганди. Відповідно, множина маркерів пропаганди прийме вигляд 2.4:

ММ = {*«Емоційність тексту», «Булінг», «Страх», «Мова ворожнечі»*. (2.4)

Таким чином, розроблена модель семантичної структури пропаганди охоплює клас тексту за рівнем пропаганди, загальну відсоткову оцінку наявності пропаганди, виявлені об'єкти пропаганди, прийоми пропаганди з їх числовими оцінками, маркери з їх числовими оцінками, їх взаємозв'язки, а також допоміжні метадані.

2.4. Взаємозв'язок методів для виявлення та класифікації прийомів та об'єктів пропаганди

Деталізована схема кроків підходу до виявлення та класифікації прийомів пропаганди наведена нижче (рис. 2.2). Вхідними даними запропонованого підходу є текстове повідомлення для аналізу, модель машинного навчання для виявлення пропаганди, множина навчених моделей машинного навчання для виявлення прийомів пропаганди та множина навчених моделей машинного навчання до оцінки сили прояву кожного маркера з доповненої множини маркерів.

Запропоновано підхід до нейромережевого виявлення [117, 118] і класифікації прийомів та об'єктів пропаганди у текстовому контенті, що складається з трьох послідовних етапів та забезпечує виявлення наявності пропаганди, класифікацію використаних прийомів пропаганди та встановлення об'єктів виявленого пропагандистського впливу.

Крок 1 – класифікації текстів за вмістом пропаганди нейромережевими моделями [119] глибокого навчання – дозволяє виявляти як явні, так і приховані пропагандистські меседжі [120, 121], що забезпечує більш глибоке розуміння послідовності та контексту в текстовому контенті.

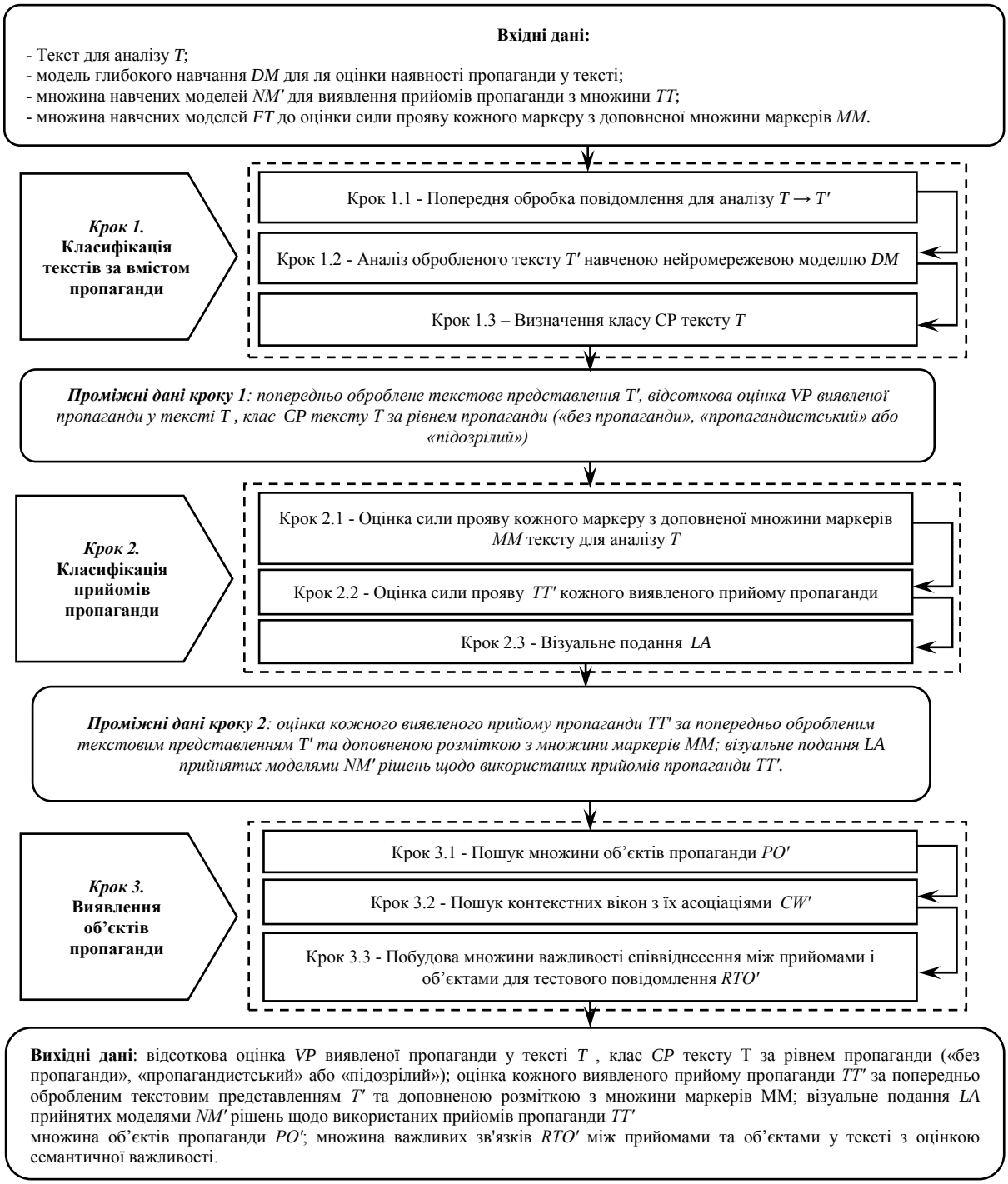


Рис. 2.2 – Деталізація підходу до виявлення та класифікації пропаганди

Крок 2 – виявлення прийомів пропаганди за маркерами [122] – дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених окремим прийомам 17-ти моделей машинного навчання у вихідні дані.

Вихідні дані містять числові оцінки наявності кожного з прийомів пропаганди із візуальною аналітикою присутності детектованих маркерів пропаганди, що забезпечує виявлення різних пропагандистських прийомів.

Крок виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень характеризується розширенням множини об’єктів пропаганди за рахунок додавання варіантів їх словесних подань і використанням контекстних вікон для виявлення взаємозв’язків між використаними прийомами та об’єктами пропаганди, що дало можливість виявляти об’єкти пропаганди та візуально їх інтерпретувати. Для візуальної інтерпретації одержаних результатів формується візуальна аналітика щодо знайдених прийомів та об’єктів пропаганди, що дозволяє візуально спостерігати об’єкти впливу в межах використовуваних прийомів пропаганди.

Вихідними даними запропонованого підходу є: відсоткова оцінка виявленої пропаганди у текстовому повідомленні; категорійна оцінка; відсоткова оцінка наявності кожного прийому пропаганди; візуальне подання прийнятих моделями глибокого навчання рішень щодо використаних прийомів пропаганди; множина об’єктів пропаганди; множина важливих зв'язків між прийомами та об’єктами у тексті з оцінкою семантичної важливості.

У результаті вирішується ряд проблем в напрямі автоматизації виявлення пропаганди, таких як відсутність комплексного аналізу взаємозв’язків прийомів і об’єктів пропаганди в текстах та відсутність узагальнень для об’єктів пропаганди і їх альтернативних згадок у текстах.

2.5. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання

2.5.1. Обмеження методу

Виявлення пропаганди у текстовому контенті в межах дисертаційного дослідження має такі обмеження: за тематичним спрямуванням; за обсягом текстової інформації; за релевантністю нейромережевих результатів у часовому континуумі; за мовою текстів.

За тематичним спрямуванням буде відбуватись виявлення пропаганди серед текстів політичної спрямованості. Такий вибір напряму обумовлений наявними навчальними даними, що більшою частиною є новинами категорії «політика». Також таке спрямування обрано через зосередження роботи навколо виявлення чорної пропаганди, яка є найбільш маніпулятивною та оманливою, має спрямованість на дискредитацію та дезінформацію. Відповідно саме чорна пропаганда має більше маркерів, завдяки яким її можна виявити засобами ШІ.

Обмеження за обсягами текстової інформації також частково пов’язане з наявними навчальними даними і в межах дослідження працює з текстовими дописами довжиною від 200 до 6300 символів. Також обмеження пов’язане з самою сутністю пропаганди. Тексти менше 200 символів не можуть повною мірою розкрити її зміст і можуть призводити до хибних результатів, а тексти понад 6300 символів є більш складними і на ряду з короткими навчальними текстами їх аналіз також буде нерелевантним.

Важливим аспектом, що також є своєрідним обмеженням, є врахування часових змін у мовних особливостях пропагандистських текстів. Оскільки дослідження базується на нейромережевих моделях, які виявляють семантичні закономірності, зміни тематичного наповнення не впливатимуть суттєво на розпізнавання загальних ознак пропаганди. Однак суттєвим фактором може

бути еволюція мовних конструкцій, сленгу та стилістичних прийомів, які можуть ускладнити виявлення пропаганди. Наведені методи виявлення пропаганди перевірялись протягом 2022 – 2025 років. За зазначений період змін у результатах не спостерігалось. Однак, навіть при суттєвій еволюції мовних конструкцій, шляхом навчання нейромереж на оновлених даних наведені методи забезпечать виявлення пропаганди, оскільки використовують нейромережеві моделі, що виявляють семантичні закономірності.

Щодо мовного обмеження – розроблений метод призначений для роботи з українською мовою. Однак якщо необхідно працювати з іншими мовами, використання методу є можливим шляхом перенавчання нейромереж на наборах даних інших мов, але з виконанням обмеження стосовно розміру вхідних даних.

2.5.2. Схема та кроки методу

Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання призначений для автоматизованої ідентифікації текстів, які містять пропагандистські елементи. Особливістю запропонованого методу є те, що він дозволяє виявляти як явні, так і приховані пропагандистські меседжі, ґрунтуючись на використанні нейромереж глибокого навчання, використанні механізму аугментації навчальних текстових даних, що дозволяє розширити кількість навчальних зразків.

Кроки методу класифікації текстів за вмістом пропаганди наведено на рис. 2.3. Метод працює шляхом перетворення вхідних даних у вигляді навченої нейромережевої моделі глибокого навчання та тексту для класифікації у вихідні дані у вигляді відсоткової оцінки наявності пропаганди у тексті та присвоєння одного з 3-х класів: «без пропаганди», «пропагандистський» та «підозрілий».

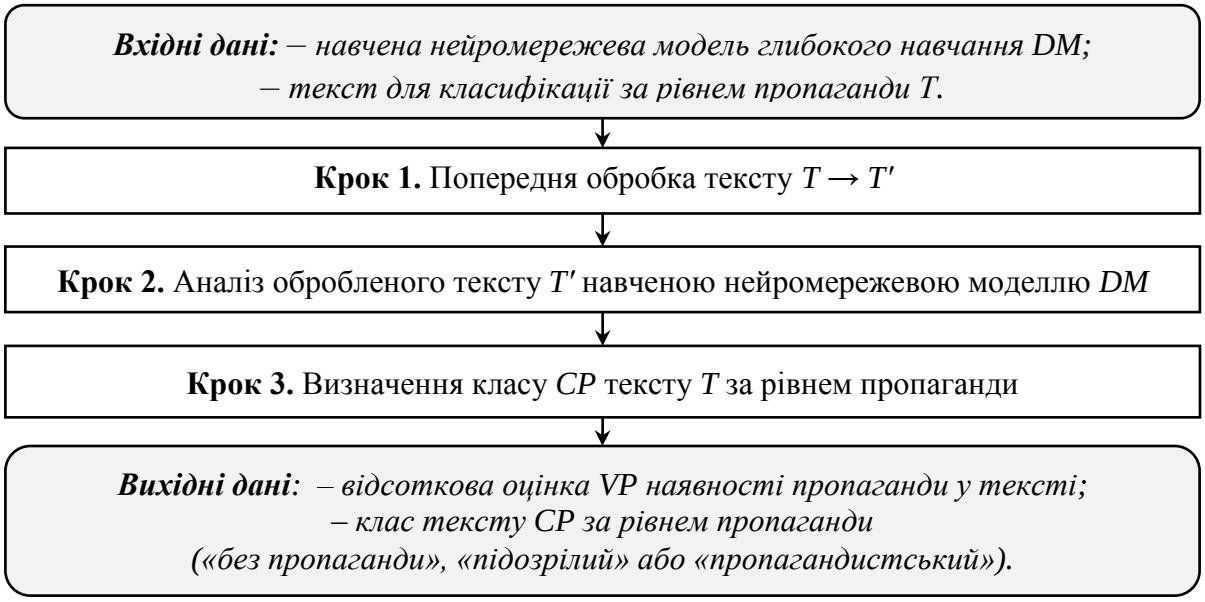


Рис. 2.3 – Кроки методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання

Першим кроком є попередня обробка тексту T для класифікації, що містить ряд кроків препроцесінгу $T \rightarrow T'$, таких як перетворення тексту у нижній регістр, видалення стоп-слів та елементів пунктуації тощо. Попередньо оброблений текст T' перетворюється у числові послідовності, які будуть подані навченій нейромережевій моделі глибокого навчання для подальшого одержання відсоткових показників наявності пропаганди.

Другим кроком є аналіз тексту навченою нейромережевою моделлю, який виконає числову оцінку рівня наявності пропаганди у тексті. Для отримання даної числової оцінки використовується нейронна мережа глибокого навчання, яка видає оцінку в діапазоні від 0 до 1, де 1 – максимально проявлена пропаганда, а 0 – не виявлена.

Кроком 3 є віднесення проаналізованого тексту за оцінкою, отриманою на кроці 2 до одного з 3-х класів: «без пропаганди», «пропагандистський» та «підозрілий». Для цього емпіричним шляхом були встановлені межі для кожного з класів. Клас «без пропаганди» має межі від 0 до 0.45, «підозрілий»

має межі від 0.45 до 0.55, та «пропагандистський» від 0.55 до 1. Межі можуть змінюватись та налаштовуватись залежно від видів пропаганди та специфіки користувацьких даних.

Відповідно вихідними даними є відсоткова оцінка наявності пропаганди у тексті та присвоєння одного з 3-х класів: «без пропаганди», «пропагандистський текст» та «підозрілий».

У п. 1.3 визначено, що для виявлення пропаганди можуть бути застосовані такі види нейромереж глибокого навчання, як рекурентні нейромережеві моделі та нейромережеві архітектури за типом трансформер [64]. Також ефективним є формування ансамблів із нейромереж глибокого навчання [63].

Ансамблі рекурентних нейромереж є ефективними, оскільки ці моделі мають високу варіативність передбачень через їхню чутливість до ініціалізації ваг і порядку обробки послідовності. Оскільки рекурентні моделі кодують інформацію поступово, їх кінцевий стан може значно змінюватися залежно від параметрів навчання. Це призводить до того, що ансамблювання допомагає компенсувати недоліки окремих моделей, згладжуючи шум та покращуючи узагальнення.

Натомість трансформери використовують механізм самоуваги, який дозволяє паралельно аналізувати всю послідовність одночасно, роблячи навчання більш стабільним. Оскільки трансформери не залежать від порядку надходження інформації так сильно, як RNN, їхні параметри мають меншу дисперсію, а значить, ансамблювання не дає такого значного приросту якості. Крім того, трансформери є значно обчислювально затратнішими, тому об’єднання кількох моделей у ансамбль робить їх використання непрактичним.

З точки зору статистики, ансамблювання ефективне, коли окремі моделі ансамлю демонструють різні джерела похибок, які можна скоригувати, комбінуючи передбачення моделей. У трансформерах, через їхню здатність

ефективно охоплювати глобальний контекст, внутрішні варіації значно менші, що знижує користь ансамблю.

Тому в межах дослідження для методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання буде перевірено використання рекурентних нейромереж у складі ансамблю та застосування гібридної архітектури, що дозволить інтегрувати переваги нейромереж-трансформерів та рекурентних нейромереж.

Таким чином, запропонований метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання дозволяє виявляти як явні, так і приховані пропагандистські меседжі й може використовуватись для оцінки потенційних загроз, пов’язаних із поширенням пропаганди.

2.5.3. Використання нейромережевих моделей BiLSTM та GRU для класифікації текстів за вмістом пропаганди

Метод виявлення пропаганди в інтернет-контенті нейромережами глибокого навчання, що призначений для виявлення та аналізу потенційно пропагандистського або маніпулятивного контенту (викладений у п. 2.5.2), при використанні нейромереж BiLSTM та GRU в ансамблевій формі набуде вигляду, як на рис. 2.4.

Вхідними даними методу є ансамбль навчених моделей рекурентних нейронних мереж з токенізаторами, і текст для аналізу (*T*). Вихідними даними є рівень і відсоткова оцінка наявності пропаганди як за кожною RNN-моделлю з використанням ансамблевого підходу, так і узагальнено.

На кроці 1 відбувається попередня обробка тексту, яка включає в себе послідовне виконання підкроків 1.1 – 1.4.

На кроці 1.1 здійснюється вибір і завантаження ансамблю RNN-моделей для визначення пропаганди, а також їх токенізаторів (крок 1.2).

Рис. 2.4 – Деталізація кроків методу класифікації текстів за вмістом пропаганди з використанням нейромережевих RNN-моделей BiLSTM та GRU

Крок 1.3 виконує попередню обробку користувацького тексту T для аналізу, що включає в себе перетворення тексту у нижній регістр, видалення стоп-слів та елементів пунктуації.

На кроці 1.4 попередньо оброблений текст перетворюється у числові послідовності за допомогою токенізаторів TKN_i, які будуть подані нейронним мережам DM_i на вхід для подальшої бінарної класифікації.

Кроком 2 є аналіз допису на наявність пропаганди, що включає в себе одержання відсоткових показників наявності пропаганди в обробленому тексті T' за аналізом кожною RNN-моделлю [117].

На кроці 3 здійснюється формування висновку щодо наявності пропаганди. Для цього пропонується використати два підходи – бінарний та дискретний. Для бінарного підходу для визначення рівня пропаганди для нейромереж ансамблю отримуються бінарні оцінки, де оцінка 0 – не містить пропаганди, 1 – містить пропаганду. У дискретному підході оцінка нейромереж береться як дискретна величина з проміжку від 0 до 1, де 1 – максимальний прояв пропаганди, а 0 – її відсутність.

У випадку бінарного підходу відбувається одержання бінарної оцінки й висновок щодо класу (*CP*) тексту *T* формується за правилами:

– «пропагандистський», якщо більше 50% нейромережевих моделей отримали бінарну оцінку «1»;

– «без пропаганди», якщо більше 50% нейромережевих моделей отримали бінарну оцінку «0»;

– «підозрілий», якщо нейромережі мають паритетні результати голосування (половина з оцінками «0» і половини з оцінками «1»).

Визначення рівня пропаганди у випадку дискретної оцінки відбувається таким чином. експертним шляхом установлюються межі трьох класів – верхня оціночна дискретна межа класу «без пропаганди» та нижня оціночна дискретна межа класу «пропагандистський», після чого здійснюється розрахунок загальної дискретної оцінки приналежності допису до вказаних класів (2.5):

$$Eval = k_1 \square RNN_1 + k_2 \square RNN_2 + \ldots + k_n \square RNN_n, \qquad (2.5)$$

де k_1, k_2, .. , k_n – коефіцієнти впливу дискретних оцінок, отриманих нейромережами RNN_1, RNN_2,.., RNN_n відповідно. У поданні 2.5 RNN_i відповідає DM_i, а оцінка *Eval*100%* відповідає загальній відсотковій оцінці наявності пропаганди *VP*.

Відповідно до вищевикладеного матеріалу, результатом роботи запропонованого методу є рівень і відсоткова оцінка наявності пропаганди за кожною RNN-моделлю ансамбля, а також узагальнений рівень і відсоткова оцінка наявності пропаганди у досліджуваному дописі.

Як вже було вище зазначено, було обрано до використання рекурентні нейронні мережі, а саме – архітектури BiLSTM та GRU. Архітектури нейронних мереж BiLSTM та GRU наведені на рис. 2.5 (а) та рис.2.5 (б).

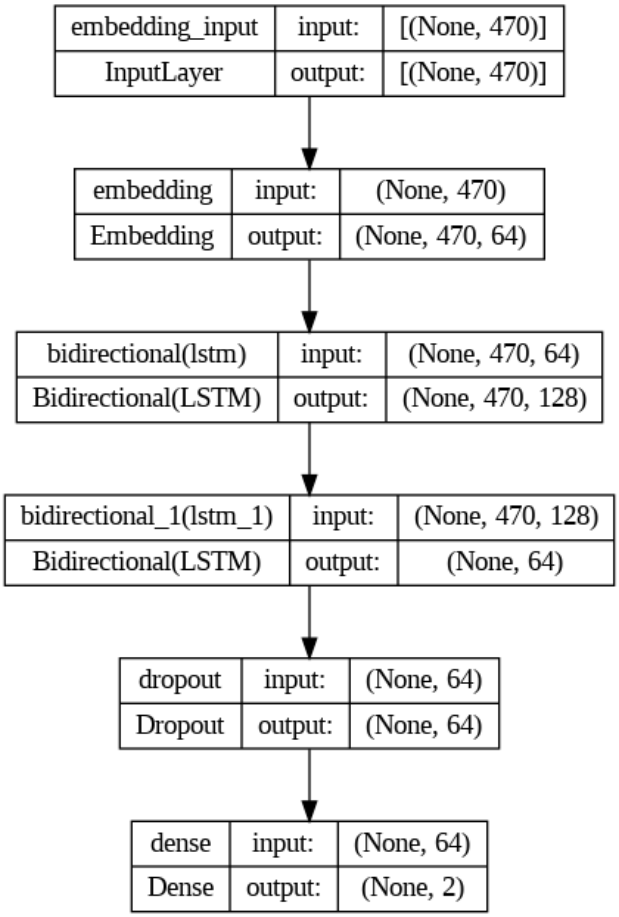


Рис. 2.5 (а) – Архітектура нейромережі BiLSTM для виявлення пропаганди

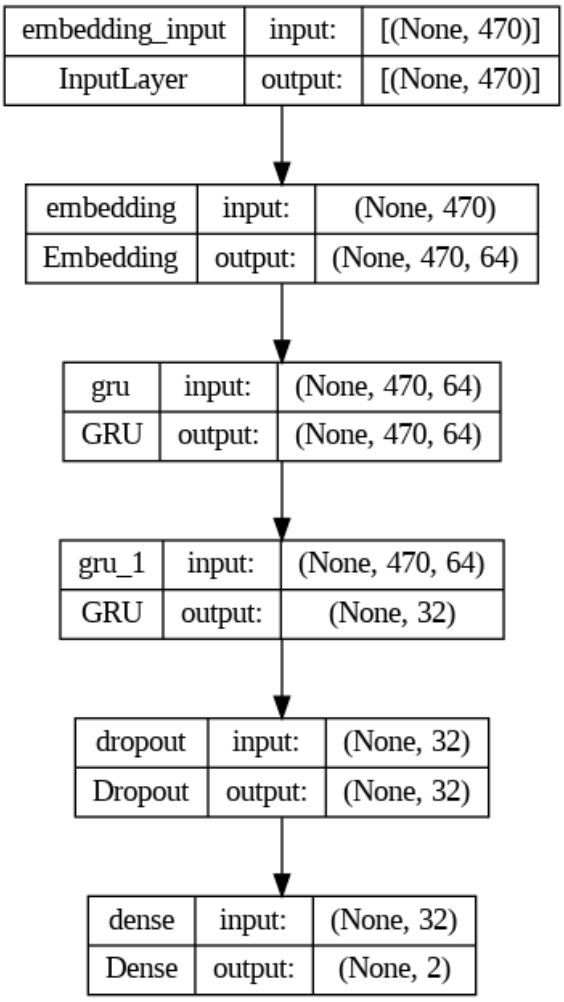


Рис. 2.5 (б) – Архітектура нейромережі GRU для виявлення пропаганди

Підбір різних нейромережевих моделей обумовлений їх специфічними можливостями щодо аналізу текстових послідовностей. BiLSTM шляхом використання прихованого стану дозволяє аналізувати текстові послідовності у прямому та зворотному напрямках, що допомагає усунути бар'єри традиційних RNN [126]. GRU має механізми воріт, які дозволяють ефективніше управляти градієнтами в часі, що робить модель більш стійкою до проблеми зникнення градієнтів порівняно з класичними RNN [127].

Приклад використання нейромережевих моделей BiLSTM та GRU для класифікації текстів за вмістом пропаганди наведено нижче.

Вхідний текст *T* має такий вигляд: «*Наші вороги знову наступають! Ми повинні негайно об'єднатися проти них, інакше нас чекає катастрофа!*».

Тоді на кроці 1 текст *T* буде попередньо оброблений та набуде таких змін, як перетворення у нижній регістр: «*наші вороги знову наступають! ми повинні негайно об'єднатися проти них, інакше нас чекає катастрофа!*». Видалення стоп-слів: «*вороги знову наступають! повинні негайно об'єднатися проти, інакше чекає катастрофа!*». Видалення пунктуації: «*вороги знову наступають повинні негайно обєднатися проти інакше чекає катастрофа*». Далі текст після попередньої обробки буде токенізований відповідними завантаженими токенізаторами TKN_1 – для BiLSTM та TKN_2 – для GRU. Заповнення нульовими значеннями спереду до довжини, якої не вистачає до 470, заповнення числовими значеннями, якщо слова є у словнику. Відповідно для BiLSTM процес попередньої обробки завершиться утвореним числовим представленням: [*0, 0, .. , 0, 37, 1854, 1, 571, 2157, 1, 1560, 1*], а для GRU [*0, 0, .. , 0, 37, 1570, 10, 468, 2067, 35, 890, 55*].

На кроці 2 здійснюється аналіз тексту RNN-моделями:

DM_1 (BiLSTM) дасть для досліджуваного попередньо обробленого тексту T' вихід 0.87 (дискретна оцінка), а бінарна оцінка при цьому буде «1».

DM_2 (GRU) дасть для досліджуваного попередньо обробленого тексту *T′* вихід 0.91, бінарна оцінка при цьому буде також «1». Відповідно відсоток наявності пропаганди за DM_1 = 87%, а відсоток за DM2 = 91%.

На кроці 3 здійснюється формування висновку. Для бінарного підходу дві моделі (DM1 та DM2) дали однакові оцінки – «1». Отже, 100% моделей дають оцінку «1». Висновок: Текст – «Пропагандистський».

Відповідно до дискретного підходу буде обраховано оцінку: *Eval* = 0.5 * 0.87 + 0.5 * 0.91 = 0.89. Межі класів: ≤0.4 – «Без пропаганди», 0.4–0.6 – «Підозрілий», 0.6 – «Пропагандистський». Висновок: *CP* = «Пропагандистський», *VP* = 89%. відсоток наявності пропаганди за DM_1 = 87%, а відсоток за DM2 = 91%.

Отже, наведена нейромережева архітектура BiLSTM для методу виявлення пропаганди у текстовому контенті на основі використання ансамблю (завдяки використанню прихованого стану) дозволяє обробляти текст у прямому та зворотному напрямках, що усуває обмеження традиційних рекурентних нейронних мереж. Це дозволяє моделі захоплювати більше контексту і покращує точність аналізу. Водночас GRU, оснащена механізмами воріт, ефективніше керує градієнтами в часі, що робить її більш стійкою до проблеми зникнення градієнтів порівняно з класичними RNN. Це сприяє кращому утриманню довгострокових залежностей у тексті, що є критично важливим для виявлення пропаганди.

2.5.4. Використання нейромережевої моделі BiLSTM гібридної архітектури для класифікації текстів за вмістом пропаганди

Метод виявлення пропаганди в інтернет-контенті нейромережами глибокого навчання, що викладений у п. 2.5.2, при використанні нейромережі BiLSTM гібридної архітектури набуде вигляду, як на рис. 2.6.

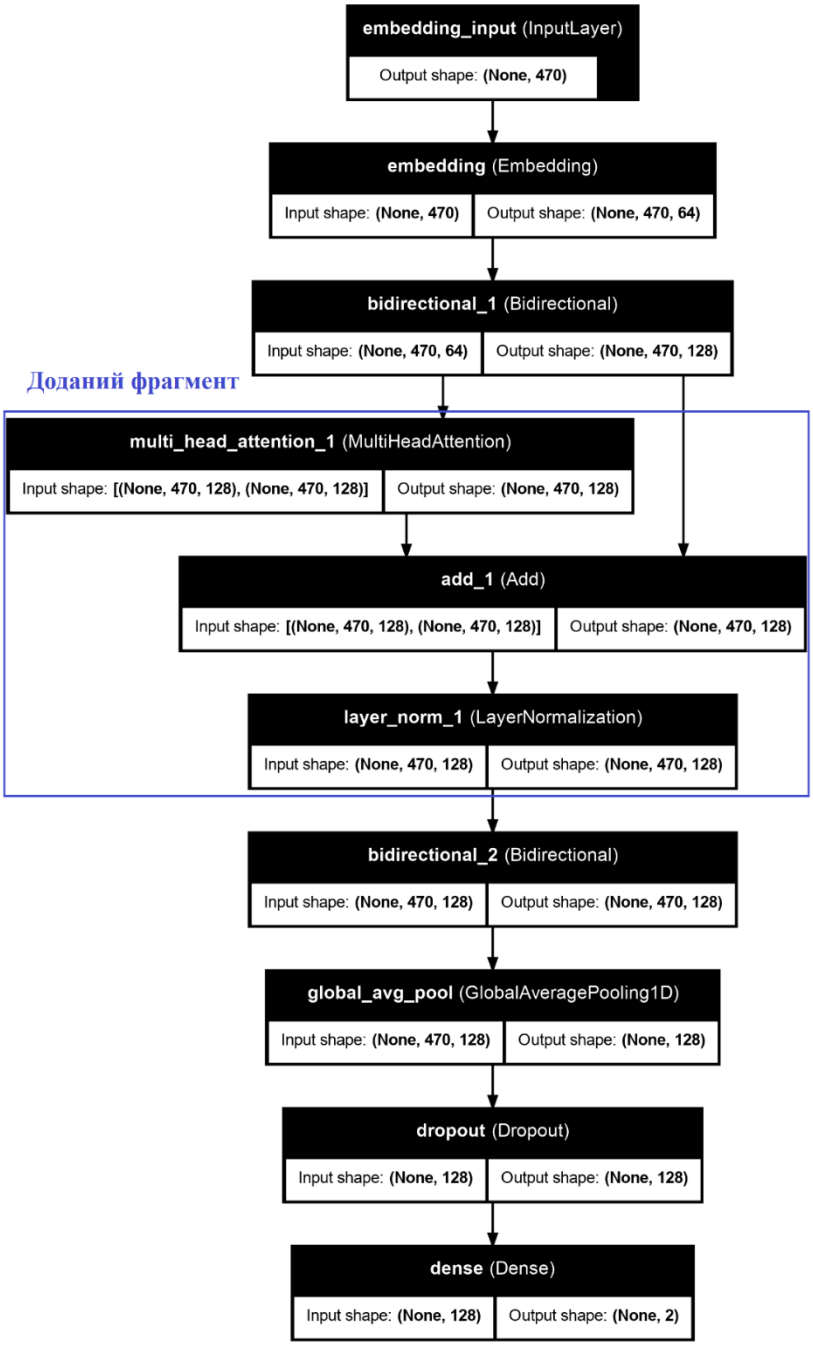


Рис. 2.6 – Гібридна архітектура нейронної мережі глибокого навчання для класифікації пропаганди

Відповідно вхідними даними є навчена нейромережева модель гібридної архітектури (як модель глибокого навчання) та текст для класифікації.

Під гібридною архітектурою в межах методу слід розуміти додавання до рекурентної архітектури шару багатоголовкової уваги. Рекурентні моделі,

навіть у двонаправленому вигляді, обробляють текст поступово, що може призводити до втрати важливих глобальних залежностей через згасання градієнтів або обмежену місткість пам’яті. Багатоголовкова увага дозволяє моделі розподіляти вагу між різними частинами тексту незалежно від їхньої відстані, забезпечуючи паралельне виявлення значущих патернів. Це покращує здатність мережі до узагальнення, оскільки вона вчиться акцентувати увагу на ключових аспектах вхідних даних, не покладаючись виключно на послідовну динаміку оновлення станів.

Що ж стосується використання моделей архітектури трансформер у чистому вигляді, то варто враховувати деякі особливості. Зокрема, трансформери навчаються на величезних корпусах, що дозволяє їм виявляти загальні мовні закономірності, однак для ефективного донавчання на специфічному завданні їм необхідно мати достатньо репрезентативну вибірку. Відповідно, враховуючи наявні дані дослідження, описані в п. 4.2, це може створювати ризик адаптації моделі до локальних закономірностей вибірки, що знижує її узагальнюючу здатність та підвищує ймовірність перенавчання.

На відміну від трансформера, BiLSTM із механізмом багатоголовкової уваги є менш чутливим до обмеженого розміру корпусу, оскільки рекурентна структура дозволяє ефективно використовувати навіть невелику кількість прикладів для формування узагальнених репрезентацій тексту. Додавання механізму уваги забезпечує додаткову гнучкість у визначенні релевантних фрагментів тексту, що робить модель більш стабільною за умов обмеженої навчальної вибірки. Таким чином, навчання гібридної архітектури з нуля є більш придатним для класифікації пропаганди, оскільки воно дозволяє зберегти контекстну інформацію без ризику втрати узагальнюючої здатності. Водночас навчання трансформеру з нуля є неефективним через обмеженість навчальних даних, а тонке налаштування може породжувати ризик адаптації моделі до локальних закономірностей вибірки.

Кроки 1-3 відбуваються аналогічно описаній у п. 2.5.2 послідовності: попередня обробка ($T \rightarrow T'$), аналіз тексту нейромережею гібридної архітектури (*DM*) та крок віднесення проаналізованого тексту *T* до одного із 3-х класів: «без пропаганди», «пропагандистський» та «підозрілий».

Вихідними даними буде, відповідно до п. 2.5.2, відсоткова оцінка наявності пропаганди у тексті *T* та присвоєння одного з 3-х класів (*CP*): «без пропаганди», «пропагандистський» та «підозрілий».

При виборі архітектур для модифікації варто зазначити, що GRU і BiLSTM є варіантами рекурентних нейромереж, які використовуються для обробки послідовностей, але вони мають фундаментальні відмінності у способі збереження контексту. GRU має спрощену архітектуру з меншим числом параметрів, що зменшує його здатність запам’ятовувати далекі залежності. Водночас BiLSTM використовує окремі потоки для прямого та зворотного проходу, що дозволяє йому зберігати більше інформації про контекст.

Багатоголовкова увага (MHA) найбільш ефективно працює у поєднанні з BiLSTM, оскільки її механізм зважування значущості елементів послідовності доповнює двонаправлене зчитування. BiLSTM кожен токен вже має контекст із майбутніх і попередніх слів, а MHA допомагає додатково підсилити релевантні зв’язки між далекими елементами. Це важливо для виявлення прихованих патернів у текстах, як-от пропаганда, де маніпулятивні прийоми можуть бути розподілені нерівномірно.

Використовувана для аналізу тексту та визначення відсоткової оцінки наявності пропаганди у тексті нейромережева архітектура (рис. 2.6) поєднує можливості рекурентних нейронних мереж та механізму уваги, забезпечуючи здатність виявляти як локальні, так і глобальні залежності в текстових послідовностях.

Спочатку вхідні токени перетворюються у векторні представлення через шар вбудовування, що дозволяє моделі працювати з щільнішими

представленнями слів. Далі BiLSTM фіксує контекстну інформацію з обох напрямків, що особливо корисно для аналізу складних текстових структур.

Однак LSTM, навіть у двонаправленій конфігурації, обмежений в ефективному захопленні далекозалежних зв’язків через лінійний характер обробки послідовностей. Саме тому на цьому етапі вводиться механізм багатоголовкової уваги, який, на відміну від рекурентних шарів, дозволяє виявляти зв’язки між токенами незалежно від їхньої відстані у послідовності. Це підвищує здатність моделі фокусуватися на релевантних частинах тексту, що особливо важливо у задачах класифікації, зокрема у виявленні пропагандистських наративів, де ключові патерни можуть бути розкидані по всьому тексту.

Після шару уваги застосовується залишкове додавання, що забезпечує стабільність градієнтів, а також нормалізація, яка прискорює навчання. Другий BiLSTM додатково зміцнює контекстне розуміння, зберігаючи послідовність обробки, а глобальний пулінг та щільний шар формують остаточне передбачення. Розташування механізму уваги між двома BiLSTM рівнями є виправданим, оскільки саме на цьому етапі модель має достатньо насичене представлення тексту, яке увага може збагачувати, тоді як застосування її до початкових ембеддингів або перед остаточними класифікаційними шарами було б менш ефективним.

Оскільки є проблема з недостатністю даних для навчання нейромережевих моделей, у межах методу використано метод аугментації тексту [128] для збільшення різноманітності навчальних даних. Для цього буде використано модель T5 (Text-to-Text Transfer Transformer) [129], яка є трансформерною нейромережевою моделлю, що обробляє всі завдання як завдання перетворення тексту в текст: переклад, узагальнення, питання-відповідь, перефразування. Використання аугментації досліджено в п. 4.2.

Використання механізму аугментації разом із багатоголовковою увагою є доцільним, оскільки увага дозволяє моделі ефективно інтегрувати зовнішні або модифіковані патерни, які виникають під час аугментованих варіацій даних.

Багатоголовкова увага здатна адаптивно розподіляти ваги між різними частинами послідовності, а тому, на відміну від традиційних рекурентних мереж, зокрема BiLSTM чи GRU, вона краще узагальнює змінені або синтетично доповнені дані. У класичних рекурентних мережах обробка послідовностей є чітко послідовною, а отже, вони менш гнучко реагують на структурні варіації вхідних даних. Крім того, довготривалі залежності в LSTM та GRU вже обмежені через поступову втрату інформації при збільшенні довжини послідовності, що робить вплив аугментації менш значущим. Натомість увага може миттєво знаходити критичні елементи навіть у зміненому контексті, що підвищує стійкість до варіацій даних і сприяє генералізації, особливо у завданнях, де пропагандистські патерни можуть мати різні форми, але зберігати спільні семантичні залежності.

Запропонований в п. 2.5.2 метод з використанням нейромережі гібридної архітектури є обґрунтованим та потребує апробації експериментальним шляхом, (наведено в п. 4.2).

Отже, в методі класифікації текстів за вмістом пропаганди пропонується використання гібридної нейромережевої моделі глибокого навчання, яка поєднує в собі традиційну рекурентну нейромережеву модель з довгостроковою пам’яттю із трансформерами, що забезпечують більш глибоке розуміння послідовності та контексту в текстовому контенті. Модифікація полягає у додаванні до базової нейромережевої моделі глибокого навчання шару багатоголовкової уваги, характерного для нейромереж архітектури трансформер. Висувається припущення, що така модифікація дозволить виявляти пропаганду з вищою точністю, ніж існуючі аналоги.

2.6. Критерії оцінювання методів виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті

Оцінювання методів виявлення та класифікації пропагандистських прийомів і об’єктів у текстовому контенті ґрунтується на комплексному підході до аналізу їхньої точності та адаптивності до змінних мовних конструкцій. Одним із ключових аспектів є здатність моделей коректно ідентифікувати пропагандистські прийоми, що вимагає оцінки їхньої точності, повноти та загальної збалансованості між цими параметрами. Використання статистичних метрик дозволяє визначити рівень відповідності прогнозованих результатів реальним прикладам, забезпечуючи об’єктивний підхід до верифікації.

Основні параметри, за яким оцінюватиметься якість класифікації, має зосередження навколо матриці помилок, яка містить такі параметри: кількість правильно передбачених позитивних випадків *TP*; кількість правильно передбачених негативних випадків *TN*; кількість хибно передбачених позитивних випадків *FP*; кількість хибно передбачених негативних випадків *FN*. За цими параметрами будуть обраховані наступні показники.

Accuracy є часткою коректно класифікованих прийомів пропаганди від загальної кількості зразків. Обраховується за формулою (2.6):

$$Accuracy = (TP+TN)/(TP+TN+FP+FN). \quad (2.6)$$

Значення Accuracy лежить у діапазоні [0, 1], де 1 означає ідеальну точність, а 0 – повну відсутність правильних передбачень.

Precision є часткою правильних позитивних передбачень класифікованих прийомів пропаганди серед усіх передбачених як визначений прийом пропаганди. Розраховується за формулою (2.7):

$$Precision = (TP)/(TP+FP). \quad (2.7)$$

Значення Precision лежить у діапазоні [0, 1], де 1 означає ідеальну влучність (усі передбачення прийому пропаганди правильні).

Recall (повнота) є часткою правильних позитивних передбачень щодо визначених прийомів пропаганди серед усіх реальних позитивних випадків. Розраховується за формулою (2.8):

$$Reccal = (TP)/(TP+FN). \quad (2.8)$$

Значення Recall лежить у діапазоні [0, 1], де 1 означає ідеальну повноту (всі справжні виявлення прийому виявлені коректно).

F_1 є середнім гармонійним між Precision та Recall та обчислюється за формулою (2.9):

$$F_1 = 2(Precision* Reccal)/(Precision + Reccal). \quad (2.9)$$

Для оцінок засосованих у дослідженні моделей будуть використані показники Precision, Recall та F_1 за наведеними вище формулами, а також усереднені значення за класами.

Вибраний набір метрик забезпечує всебічну оцінку якості класифікаційних моделей глибокого навчання, оскільки Accuracy [74] відображає загальну точність передбачень, тоді як Precision і Recall [117] дозволяють оцінити компроміс між хибнопозитивними та хибнонегативними передбаченнями. Використання F_1 як гармонійного середнього Precision і Recall є особливо важливим для незбалансованих наборів даних, оскільки він узгоджує точність і повноту моделі.

Отже, у дослідженні навчені моделі будуть оцінені за таким набором метрик: Accuracy, Precision, Recall, F_1 міра. Такий набір метрик є достатнім для оцінки моделей в рамках виявлення пропаганди та класифікації маніпулятивних прийомів з огляду на особливості завдання, зокрема багатокласовий характер класифікації. Оскільки кожна модель відповідає за конкретний маніпулятивний прийом, така оцінка дозволить точно визначити, де кожна з моделей працює добре, а де – потребує доопрацювання.

2.7. Висновки до розділу 2

У другому розділі дисертації запропоновано концепцію виявлення і класифікації прийомів та об’єктів пропаганди у текстовому контенті з використанням нейромережевевих засобів, що містить етапи класифікації текстів за вмістом пропаганди, виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень та виявлення об’єктів пропаганди з візуальною інтерпретацією прийнятих рішень. Запропонована концепція складається з трьох послідовних етапів: класифікації текстів за вмістом пропаганди; виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією результатів; виявлення об'єктів пропаганди з їх візуальною аналітикою.

На першому етапі здійснюється класифікація текстів за допомогою моделі глибокого навчання, що дозволяє визначити наявність або відсутність пропаганди, а також ідентифікувати підозрілі тексти. Другий етап включає аналіз текстів, які класифіковані як пропагандистські, на предмет використання різних прийомів пропаганди за допомогою 17-ти окремих нейромережевих моделей. На третьому етапі відбувається виявлення об'єктів пропаганди та їх взаємозв'язків із використаними прийомами, що дозволяє створити детальну візуальну аналітику.

Завдання виявлення та класифікації пропагандистських прийомів у текстовому контенті тісно пов’язане з політичними, соціальними й культурними контекстами, що створює підвищений ризик викривлення або хибного маркування інформації. Штучний інтелект, спираючись на навчальні дані, здатен відтворювати або навіть підсилювати існуючі людські упередження, що становить загрозу свободі слова, може сприяти маніпуляції громадською думкою або дискримінації окремих груп. Тому відзначено, що важливо забезпечувати не лише точність моделей, а й пояснюваність їхніх рішень, а

також залишати остаточну відповідальність за інтерпретацію результатів за людиною.

У межах цього дослідження ШІ виконує допоміжну функцію: його використання обмежується інструментами для виявлення та попередньої класифікації, тоді як кінцеве рішення ухвалює людина. Запропоновані методи характеризуються високою відтворюваністю завдяки детальному опису алгоритмів і даних, а також прозорістю, що досягається через візуалізацію результатів і пояснень рішень моделей.

Запропонована модель семантичної структури пропагандистського тексту використовує нейромережеві засоби для виявлення об'єктів пропаганди та їх асоціації з прийомами пропаганди. Модель дозволяє проводити глибокий аналіз текстового контенту, виявляючи складні зв'язки та контексти, що використовуються у пропагандистських матеріалах.

Метод виявлення пропаганди на основі ансамблів нейромережевих моделей спроможний демонструвати точність понад 90% у визначенні наявності пропаганди в текстах, використовуючи як бінарний, так і дискретний підходи для оцінки. Окрім того, розроблено метод на основі нейромережі гібридної архітектури, яка поєднує рекурентні нейромережі з трансформерами, що дозволяє виявляти як явні, так і приховані пропагандистські меседжі.

Отже, запропонований підхід та метод виявлення і класифікації пропагандистських прийомів у текстовому контенті демонструють і можуть бути використані для автоматизованого аналізу інформаційних загроз, сприяючи підвищенню рівня інформаційної безпеки та протидії маніпулятивним впливам на суспільство.

РОЗДІЛ 3.
МЕТОДИ ВИЯВЛЕННЯ ПРИЙОМІВ ТА ОБ'ЄКТІВ ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ

3.1. Обмеження методів виявлення прийомів та об'єктів пропаганди у текстовому контенті

При розв'язанні задач виявлення прийомів та об'єктів пропаганди у текстовому контенті виконуються обмеження, наведені в п. 2.5.1, а також додатково визначаються обмеження за кількістю прийомів пропаганди і за кількістю маркерів.

Існуюче *обмеження за кількістю прийомів*, які можна класифікувати розробленими методами, також пов'язане із джерелом даних для навчання, однак при наявності нових даних для навчання, кількість прийомів пропаганди можна розширити за допомогою запропонованих методів, які є універсальними за умови повторного навчання нейромережевих архітектур. Як було вище зазначено, в межах дисертаційного дослідження виконується класифікація за 17-ма прийомами пропаганди, визначеними у п. 1.6 та формалізованими у п. 2.3.

Щодо *обмеження за кількістю* маркерів, то на етапі дослідження їх кількість становить 4 (див. п. 2.4), однак, ця множина також може бути розширена за умови підбору або вже навченої нейромережі для додаткового маркера (за умови виконання обмежень, описаних у п. 2.5.1), або ж за умови наявності навчального набору даних й виконання процесу навчання та збереження нової нейромережевої моделі.

Вищенаведені обмеження стосуються як методу виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень, так і методу виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.

3.2. Метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень

3.2.1. Підхід до виявлення прийомів пропаганди за маркерами і візуальної інтерпретації прийнятих рішень

Комплексна схема підходу, що ілюструє усі кроки та етапи обробки даних для виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, наведена на рис. 3.1.

Представлений підхід спрямований на ідентифікацію маркерів пропаганди та аналіз пропагандистських прийомів за допомогою використання нейромережевих моделей і методів пояснюваного штучного інтелекту. Процес передбачає поетапну обробку текстових даних, навчання моделей для виявлення маркерів та оцінку сили їхнього впливу, а також навчання моделей для виявлення прийомів пропаганди.

На першому етапі здійснюється попередня обробка текстів, які використовуються для ідентифікації маркерів пропаганди. Після цього тексти розподіляються на навчальні та валідаційні вибірки, що дозволяє підготувати дані до подальшого навчання нейромережевих моделей. Основним результатом цього етапу є створення моделей, здатних розпізнавати маркери пропаганди та їхній контекстуальний вплив.

Другий етап спрямований на оцінку сили прояву маркерів. Для цього здійснюється аналіз навчальних текстів із використанням отриманих моделей, що дозволяє розширити множину навчальних даних для кожного прийому пропаганди. Це забезпечує якісніший розподіл текстів відповідно до виявлених закономірностей у контексті пропагандистських матеріалів.

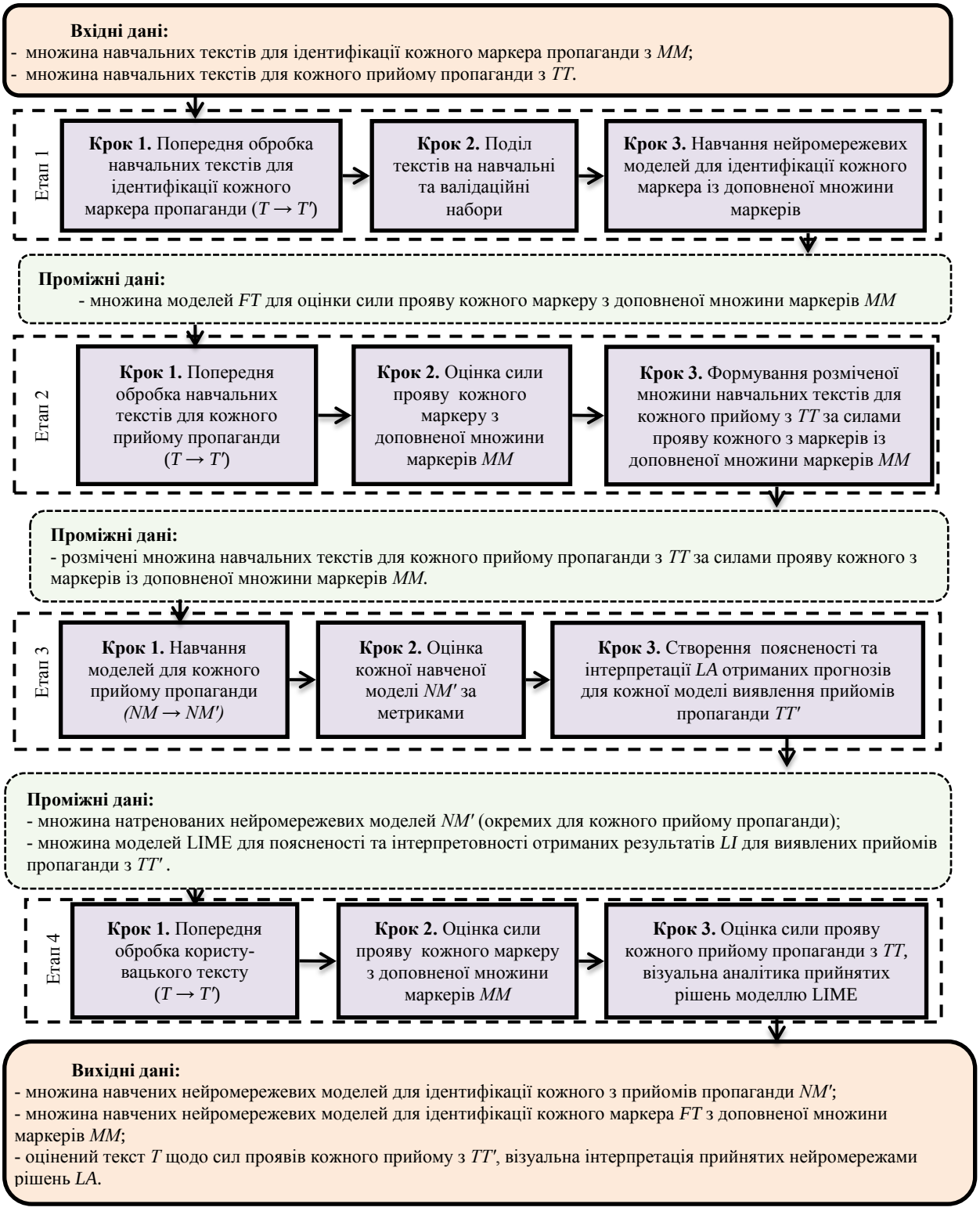


Рис. 3.1 – Комплексна схема підходу до виявлення прийомів пропаганди

На третьому етапі здійснюється процес навчання нейромережевих моделей для розпізнавання конкретних прийомів пропаганди.

Оцінка якості моделей проводиться з використанням метрик, після чого створюються пояснення та інтерпретації отриманих прогнозів.

Пояснення та інтерпретації прогнозів дозволяють сформувати комплексне уявлення про те, які маркери та прийоми домінують у текстах і які патерни використання вони демонструють.

Заключний етап передбачає практичне застосування натренованих моделей для оцінки нових текстів, зокрема аналізу користувацьких даних. Визначається сила прояву кожного маркера, після чого здійснюється пояснення отриманих висновків за допомогою методів пояснюваного штучного інтелекту, зокрема LIME. Це дозволяє не лише класифікувати тексти за наявністю пропагандистських прийомів, а й забезпечити візуальну інтерпретацію рішень моделей.

На рис. 3.2 зображено три альтернативні підходи до класифікації прийомів пропаганди: мультикласова, мультилейблова та послідовність бінарних класифікаторів, що демонструють різні способи використання нейромереж для виявлення пропагандистських прийомів.

На противагу мультикласовій класифікації іноді пропонується використовувати мультилейблову класифікацію. Перевагою такої класифікації є те, що мультилейблова класифікація враховує випадки, коли прийоми пропаганди можуть одночасно відноситися до більш ніж одного класу, і при цьому класи не є взаємовиключними. Це дозволяє уникнути проблеми втрати менш виражених прийомів пропаганди, однак залишає проблему низької точності класифікації, що є характерною як для мультикласової, так і для мультилейблової класифікації. Низька точність зумовлена відсутністю коректно сформованих навчальних даних, які б були розмічені фахівцями та містили достатню кількість зразків для навчання.

Іншою проблемою є те, що навіть за наявності розмітки, деякі прийоми пропаганди мають схожі ознаки у текстовому або візуальному контенті.

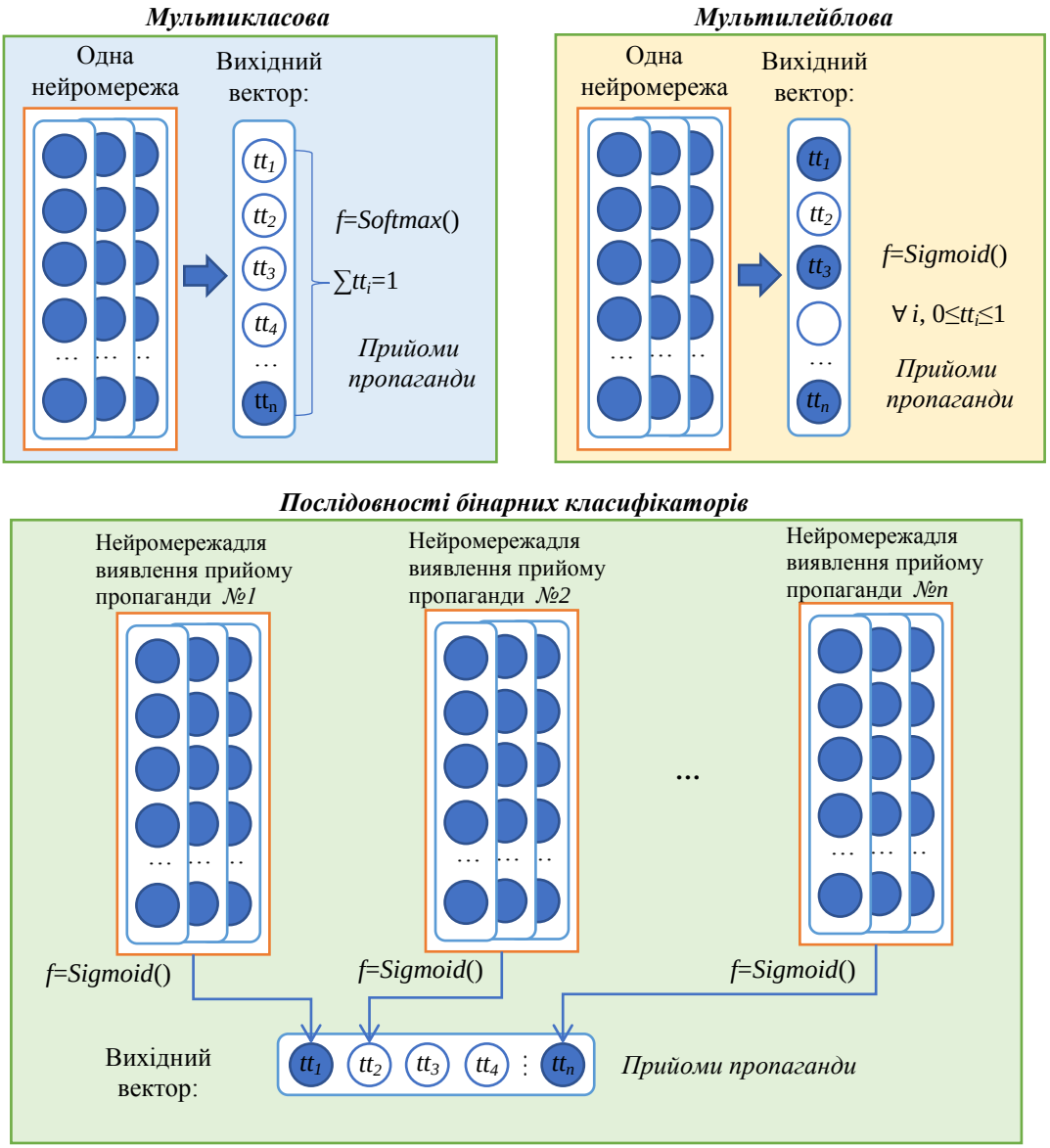


Рис. 3.2 – Підходи щодо нейромережевої класифікації прийомів пропаганди

Висувається припущення, що для підвищення точності ідентифікації прийомів пропаганди необхідно коректно формувати навчальні вибірки, які містили б цільовий та інші прийоми, а також нейтральні тексти без ознак пропаганди.

Також висувається припущення, що використання послідовності бінарних класифікаторів є більш ефективним, ніж одного мультилейблового.

Запропонований підхід враховує протиріччя, що виникає під час аналізу пропагандистських прийомів, а саме виявлення найбільш очевидних пропагандистських прийомів із втратою менш виражених, але потенційно також однаково впливових, або ж трохи менше виражених прийомів. Така ситуація може призводити до недооцінки комплексного впливу інформаційного середовища та втрачання можливостей для раннього розпізнавання маніпуляцій.

Таким чином, запропонований підхід поєднує нейромережеві технології та інтерпретовані методи аналізу, що дозволяє не тільки ідентифікувати пропагандистські маркери та прийоми, а й оцінювати їхню силу впливу та пояснювати механізми роботи моделі.

3.2.2. Особливості формування навчальних множин даних для нейромережевих моделей класифікації прийомів пропаганди

Для кожної з 17-ти типових моделей машинного навчання буде сформовано власний дочірній набір текстів, що буде задовольняти такі вимоги:

– містити тексти з визначеним прийомом пропаганди;

– у противагу використовувати набір «Інші прийоми пропаганди», доповнений текстами без пропаганди та текстами, що представляють інші прийоми пропаганди, відмінні від цільового виду.

Приклад формування набору даних [43] для виявлення прийому «Апеляція до страху» наведено на рис. 3.3.

У дослідженні буде використано 18 класів: 17 цільових, що є репрезентативними по кількості та відповідають 17-ти визначеним прийомам пропаганди та 5 – об’єднаних в категорію «Інші прийоми пропаганди».

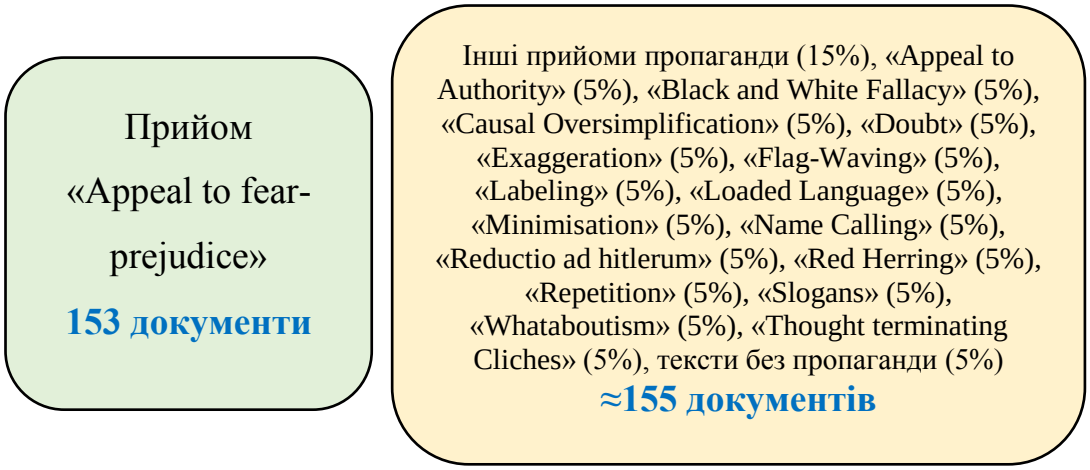


Рис. 3.3 – Приклад балансування при формуванні набору даних для навчання та тестування моделі виявлення прийому «Апеляція до страху»

Формування дочірніх наборів текстів для кожної з 17-ти типових моделей машинного навчання, зокрема шляхом балансування цільових і некатегоризованих текстів, має свої переваги.

Наведений підхід до балансування даних при формуванні набору даних для навчання та тестування моделі виявлення прийомів дозволяє чіткіше ідентифікувати конкретні прийоми пропаганди, покращуючи точність моделей. Використання набору «Інші прийоми пропаганди» (як контр приклад) зменшує ймовірність змішування різних прийомів пропаганди, що підвищує специфічність класифікації [130, 131]. Також використаний підхід до формування навчальних множин даних для нейромережевих моделей класифікації прийомів пропаганди сприяє здатності моделі до узагальнення, оскільки вона вчиться розрізняти пропагандистські прийоми в широкому контексті, враховуючи варіативність реальних текстів. Це підвищує надійність виявлення пропаганди в різноманітних джерелах інформації, забезпечуючи більш комплексну оцінку та покращуючи інструменти інформаційної безпеки.

3.2.3. Схема та кроки методу

Метод виявлення прийомів пропаганди за маркерами призначений для оцінки текстового контенту на предмет наявності прийомів пропаганди та визначення сили їх проявів. Метод відрізняється від існуючих тим, що враховує при навчанні нейромережевих класифікаторів додаткову множину маркерів та дозволяє здійснювати візуальну інтерпретацію отриманих результатів.

Під додатковою множиною семантичних маркерів слід розуміти використання різноманітних текстових ознак, які притаманні визначеним прийомам пропаганди. У таблиці 3.1 наведено приклад інтенсивності прояву семантичних маркерів, таких як: «Емоційність тексту» [74], «Булінг» [65], «Страх», «Мова ворожнечі» у прийомах пропаганди.

Прийом пропаганди «Апеляція до страху» використовує потенційні загрози або небезпеки для впливу на поведінку аудиторії. Його завдання полягає в тому, щоб викликати відчуття тривоги чи загрози з метою нав'язати певне рішення або точку зору. Такий прийом супроводжується високим рівнем емоційності, оскільки звернення до базових інстинктів виживання завжди активує сильні емоції. Через домінування мотиву загрози проявляється високий рівень булінгу, який формується за рахунок тиску й примусу. Висока інтенсивність страху є центральною ознакою цього прийому, тоді як мова ворожнечі проявляється слабо або нейтрально, оскільки загроза може бути узагальненою, без спрямування проти конкретної групи.

«Причинно-надмірна спрощеність» полягає в поданні складних соціальних чи політичних явищ у надто спрощеній формі, де складні взаємозв’язки зведені до однієї простої причини. Така риторика не апелює до емоцій напряму, тому рівень емоційності та булінгу є нейтральним. Прийом не викликає відчуття страху та не апелює до ворожнечі, натомість створює ілюзію

ясності й раціональності. Відтак усі прояви маркерів у цьому випадку лишаються або нейтральними, або низькими.

Таблиця 3.1

Сила проявів маркерів для прийомів пропаганди

Прийом пропаганди	**Емоційність тексту**	**Булінг**	**Страх**	**Мова ворожнечі**
«Апеляція до страху»	Висока	Висока	Висока	Нейтрально
«Причинно-надмірна спрощеність»	Нейтрально	Нейтрально	Нейтрально	Низька
«Сумнів»	Нейтрально	Нейтрально	Висока	Нейтрально
«Перебільшення»	Висока	Нейтрально	Нейтрально	Нейтрально
«Навішування ярликів»	Нейтрально	Нейтрально	Нейтрально	Висока
«Заряджена мова»	Висока	Нейтрально	Нейтрально	Висока
«Мінімізація»	Нейтрально	Нейтрально	Нейтрально	Низька
«Обзивання»	Нейтрально	Висока	Нейтрально	Висока
«Зведення до Гітлера»	Висока	Нейтрально	Висока	Висока
«Ватабоутізм»	Нейтрально	Нейтрально	Нейтрально	Висока

«Сумнів» – це метод впливу, при якому піддаються критиці достовірні факти або рішення з метою породження невпевненості. Він має нейтральний рівень емоційності, оскільки зазвичай вдається до аналітичної, логічної мови. Водночас маркер страху проявляється високо, адже невизначеність, сумнів і нестабільність у поданні інформації створюють атмосферу тривоги. Булінг і мова ворожнечі при цьому не є домінантними.

«Перебільшення» – це використання гіперболізації певних фактів або явищ для посилення емоційного ефекту. Такий прийом характеризується високою емоційністю, оскільки маніпулювання масштабами можливих подій або явищ, викликає сильні відгуки у людській свідомості. Водночас інші маркери, як-от, страх булінг і мова ворожнечі не є ключовими й залишаються на нейтральному рівні.

«Навішування ярликів» передбачає використання стереотипних або упереджених формулювань для знецінення особи або групи. Цей прийом виявляє себе через мову ворожнечі, інтенсивність якої є високою, адже «ярлик» несе в собі соціальну стигму. Інші маркери – емоційність, булінг і страх – лишаються нейтральними, оскільки основна дія відбувається на рівні позначення, а не прямого емоційного впливу.

«Заряджена мова» – це вживання емоційно насичених слів і виразів для виклику певного ставлення або реакції. Такий підхід викликає високий рівень емоційності і супроводжується активним використанням мови ворожнечі. Булінг і страх, однак, можуть бути відсутніми або виявлятися помірно, оскільки головна мета – не залякати, а вплинути через сильну емоційну риторику.

«Мінімізація» полягає у применшенні серйозності проблеми або її наслідків. Вона не передбачає відкритого емоційного тиску, тому емоційність, булінг і страх мають нейтральний прояв. Мова ворожнечі також не характерна для цього прийому. Її інтенсивність оцінюється як низька, оскільки такий прийом частіше функціонує як форма іронії або уникання відповідальності.

«Обзивання» – це пряме використання образливих слів або висловів щодо певної особи чи групи. Через природу висловлювань виявляється високий рівень булінгу та мови ворожнечі. Емоційність може бути нейтральною, адже часто цей прийом використовується без насиченого контексту. Маркер страху не є центральним, оскільки дія спрямована на приниження, а не на залякування.

«Зведення до Гітлера» – риторичний прийом, який полягає у порівнянні опонента або його поглядів із нацистською ідеологією, що спричиняє ефект дискредитації. Це призводить до високої емоційності та страху через звернення до історичної травми. Високою є також мова ворожнечі, адже здійснюється стигматизація. Булінг при цьому проявляється менш явно або нейтрально.

«Ватабоутізм» – тактика, за якої на критику відповідають вказівкою на інші проблеми, щоб уникнути відповідальності. Хоча емоційність і страх залишаються нейтральними, цей прийом часто супроводжується високим рівнем мови ворожнечі, оскільки за його допомогою відбувається відведення уваги від суті проблеми шляхом агресивного порівняння або обвинувачення.

При виявленні прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень забезпечується перетворення вхідних даних у вигляді тексту для аналізу та навчених окремим прийомам моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст із візуальною аналітикою присутності детектованих маркерів пропаганди

Схему методу виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень наведено на рис. 3.4.

Метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень виконує перетворення вхідних даних у вигляді множини навчених нейромереж *FT* для оцінки сили прояву кожного маркера з доповненої множини маркерів *MM*; множини навчених нейромереж для оцінки сили прояву кожного прийому пропаганди *NM′* та тексту для аналізу *T* у вихідні дані у вигляді оцінки проявлених прийомів пропаганди (множина *TT′*) за текстом та доповненою розміткою з множини маркерів та візуального подання прийнятих нейромережевими моделями рішень щодо наявності прийомів пропаганди [132].

Рис. 3.4 – Кроки методу виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень

Метод забезпечує виявлення 17-ти основних прийомів пропаганди: «Апеляція до авторитету», «Помилка хибної дихотомії», «Сумнів», «Причинно-надмірна спрощеність», «Перебільшення», «Розмахування прапором», «Заряджена мова», «Навішування ярликів», «Мінімізація», «Обзивання», «Зведення до Гітлера», «Червоний оселедець», «Повторення», «Гасла», «Ватабоутізм», «Кліше, що припиняють думання», «Апеляція до страху» та використовує 4 види семантичних маркерів: «емоційність тексту», «булінг», «страх» та «мова ворожнечі».

Таким чином, метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень для тестового тексту дозволяє одержувати числові оцінки наявності кожного з прийомів пропаганди та розмічений текст із візуальною аналітикою та додатковою поясненістю за детектованими семантичними маркерерами пропаганди, які показують, як різні пропагандистські прийоми пов'язані з емоційністю тексту, булінгом, страхом і мовою ворожнечі.

3.2.4. Деталізація процесу виявлення прийомів пропаганди

Оскільки метод виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень вхідними даними має вже навчені нейромережеві моделі *FT* для оцінки маркерів та моделей *NM′* для оцінки прийомів пропаганди, на рис. 3.5 наведено повну методологію виявлення прийомів пропаганди за маркерами – від створення моделей до оцінки тексту *T*.

Запропонована *методологія відрізняється від розробленого методу* виявлення прийомів пропаганди за маркерами тим, що відображає не тільки кроки одержання вихідних даних методу, а й кроки одержання нейромережевих моделей для розмітки текстів за семантичними маркерами, моделей для розмітки за прийомами пропаганди та інтерпретаційних моделей. Відтак методологія відображає процес перетворення вхідних даних у вигляді множини навчальних текстів для ідентифікації кожного маркера пропаганди; множини навчальних текстів – для кожного прийому пропаганди та тестового тексту – для виявлення прийомів пропаганди у вихідні дані у вигляді множини навчених нейромережевих моделей *NM′* для ідентифікації кожного з прийомів пропаганди *TT*; множини навчених нейромережевих моделей *FT* – для ідентифікації кожного маркера з доповненої множини маркерів *MM* та оціненого тексту –

щодо сил проявів кожного прийому пропаганди *TT*; візуальна інтерпретація прийнятих нейромережами рішень.

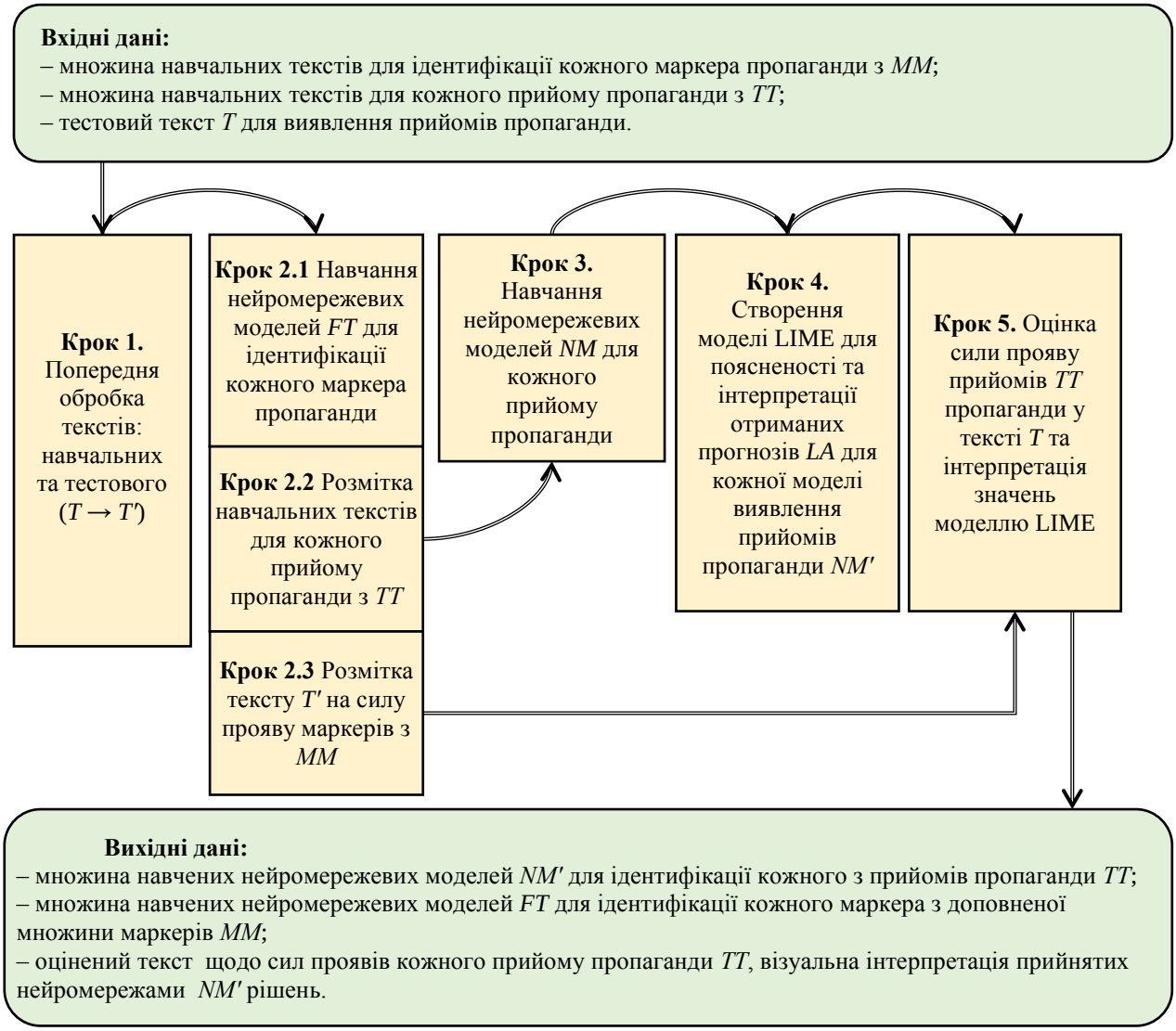


Рис. 3.5 – Методологія виявлення прийомів пропаганди за маркерами

Першим кроком є попередня обробка усіх текстових даних, як навчальних, так і текстових. Обробка включає видалення знаків пунктуації та стоп-слів, тобто кожен текст приводиться до вигляду *T′*.

На другому кроці здійснюється навчання нейромережевих моделей *FT* для ідентифікації усіх маркерів пропаганди [133], які використовуються для

розмітки навчальних текстів для кожного прийому пропаганди з множини *TT*, а також для розмітки щодо наявності маркерів тестових текстів [134].

Третім кроком є навчання нейромережевих моделей *NM* для кожного прийому пропаганди з множини *TT*. Кількість нейромережевих моделей у даному дослідженні складає 17 і покриває 17 вищеописаних прийомів пропаганди.

На четвертому кроці відбувається створення моделі LIME для поясненості та інтерпретації отриманих прогнозів для кожної моделі виявлення прийомів пропаганди, які разом із навченими нейромережевими моделями на кроці 3 будуть оцінювати користувацький текст.

На п’ятому кроці відбувається нейромережева оцінка сили прояву прийомів пропаганди у тестовому тексті *T′* та інтерпретація значень моделлю LIME [135] для формування подання *LA*.

Відповідно вихідними даними є множина навчених нейромережевих моделей *NM′* для ідентифікації кожного з прийомів пропаганди; множина навчених нейромережевих моделей *FT* – для ідентифікації кожного маркера з доповненої множини маркерів *MM* та оцінений текст – щодо сил проявів кожного прийому пропаганди з візуальною інтерпретацією *LA* щодо прийнятих нейромережами рішень.

Після використання навчених нейромережевих моделей *FT* для ідентифікації кожного маркера пропаганди, які використовуються для розмітки навчальних текстів для кожного прийому пропаганди з множини *TT*, таблиця 3.1 прийме вигляд, наведений в таблиці 3.2.

Оцінка виводиться в діапазоні від 0 до 1, де 0 – відсутність прояву семантичного маркера, а 1 – беззаперечна присутність.

Таблиця 3.2 виведена за допомогою аналізу еталонних текстів, що містять пропагандистські прийоми [134] і наводить усереднені оцінки проявів сили маркерів для прийомів пропаганди.

Таблиця 3.2

Усереднена сила проявів маркерів для прийомів пропаганди

Прийом пропаганди	**Емоційність тексту**	**Булінг**	**Страх**	**Мова ворожнечі**
«Апеляція до страху»	0.7	0.5	0.8	0.7
«Причинно-надмірна спрощеність»	0.4	0.4	0.4	0.3
«Сумнів»	0.5	0.4	0.7	0.5
«Перебільшення»	0.8	0.4	0.7	0.6
«Мітки»	0.7	0.4	0.5	0.7
«Заряджена мова»	0.8	0.4	0.5	0.8
«Мінімізація»	0.4	0.4	0.4	0.3
«Обзивання»	0.8	0.7	0.5	0.8
«Зведення до Гітлера»	0.9	0.7	0.9	0.9
«Ватабоутізм»	0.5	0.4	0.4	0.7

Таблиця 3.2 може бути використана як джерело оцінювання відповідності виявлених прийомів пропаганди і отриманих нейромережевих рішень. Це дозволить також внести додаткову поясненість щодо рішень, прийнятих ШІ, та є кроком в напрямі досягнення поставленої мети дослідження.

Схема формування нейромережевих моделей для виявлення прийомів пропаганди та моделей LIME для інтерпретованості прогнозів навчених моделей наведена на рис. 3.6.

Вхідними даними є множина навчальних текстів для кожного прийому пропаганди та доповнена множина маркерів до кожного з навчальних текстів.

Першим кроком є попередня обробка навчальних текстів, що включає видалення знаків пунктуації, зайвих пробілів та стоп-слів, тобто перетворення $T \rightarrow T'$ для кожного тексту з множини навчальних.

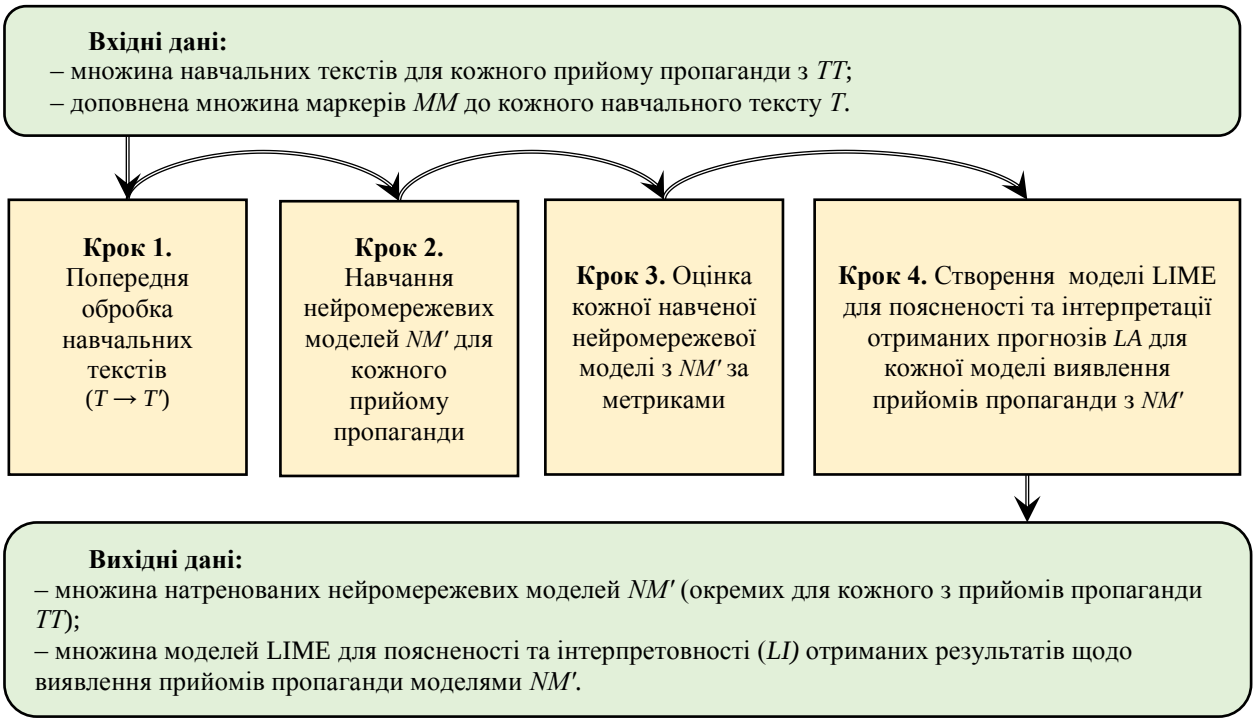


Рис. 3.6 – Процес формування нейромережевих моделей для виявлення прийомів пропаганди та моделей LIME для інтерпретованості прогнозів навчених моделей

Наступним кроком є навчання нейромережевих моделей *NM* для кожного прийому пропаганди. В якості нейромережевих моделей будуть використані BERT-подібні моделі, оскільки даний вид моделей глибокого навчання дозволяє розуміти контекст, що є важливим фактором при виявлення прийомів пропаганди.

Після процесу навчання кожна модель оцінюється за метриками точності, влучності та повноти. Для моделей, що показали найвищі значення за метриками, виконується крок 4 – створення моделі LIME для поясненості та інтерпретації отриманих прогнозів для кожної моделі виявлення прийомів пропаганди.

Модель LIME слугує для пояснення як текстових даних, так і для числових на основі проявів маркерів пропаганди, що дозволяють інтерпретувати, як модель приймає рішення на основі вхідних даних.

Вихідними даними є множина натренованих нейромережевих моделей *NM'* (окремих для кожного прийому пропаганди) та множина моделей LIME (*LI*) для поясненості та інтерпретованості отриманих результатів щодо виявлення прийомів пропаганди нейромережевими моделями.

Формування навчального набору даних для виявлення нейромережею прийомів пропаганди наведено на рисунку 3.7. Деталізований процес подання векторизованого тексту кожній моделі машинного навчання для розмітки за прийомами пропаганди наведено на рисунку 3.8.

Першим кроком є попередня обробка навчальних текстів по кожному з прийомів пропаганди, після чого відбувається другий крок – оцінка сили прояву кожного маркера з доповненої множини маркерів кожного тексту з множин навчальних текстів для виявлення прийомів пропаганди.

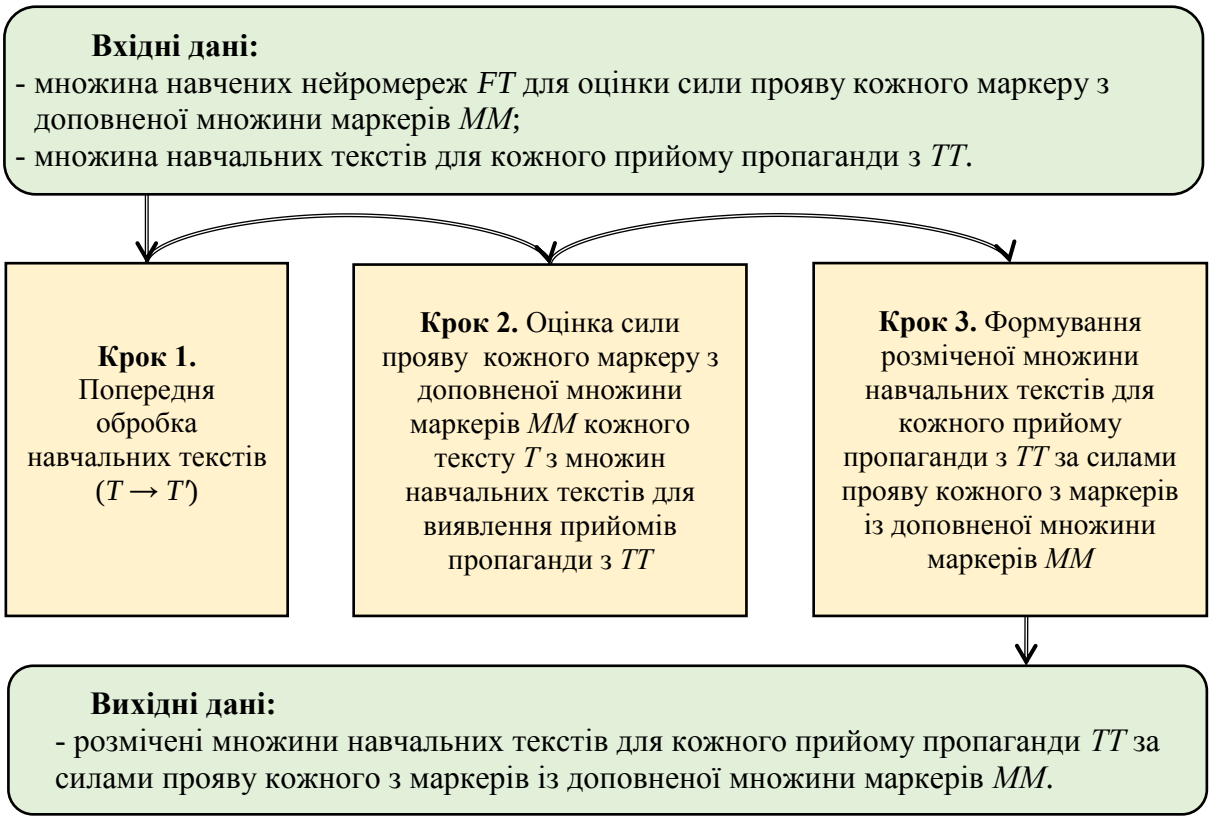


Рис. 3.7 – Процес формування навчального набору даних для нейромережевого виявлення прийомів пропаганди

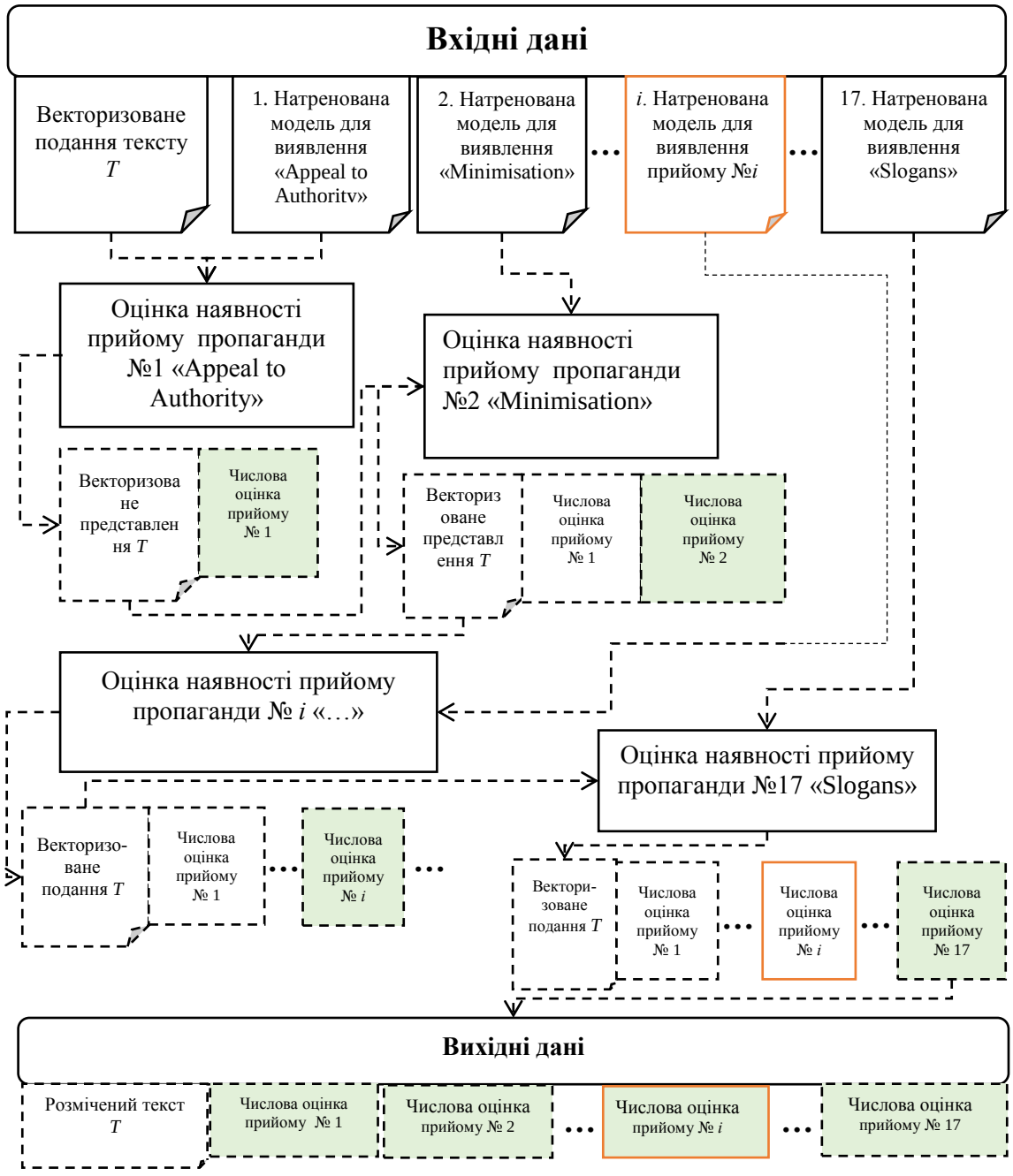


Рис 3.8 – Деталізація процесу подання векторизованого тексту кожній з моделей машинного навчання для розмітки за прийомами пропаганди

Наведене дозволяє уточнити межі застосування відповідних маркерів та підвищити точність класифікації. На основі отриманих числових оцінок здійснюється розмітка корпусу навчальних текстів, що формує стандартизований набір даних для подальшого навчання моделей.

Наступним кроком є формування розміченої множини навчальних текстів для кожного прийому пропаганди за силою прояву кожного з маркерів із доповненої множини маркерів.

Відповідно вихідними даними є розмічена множина навчальних текстів для кожного із 17-ти прийомів пропаганди є використанням сили проявів кожного з маркерів із доповненої множини маркерів. Моделі аналізують векторизовані представлення текстів і визначають рівень присутності відповідних пропагандистських прийомів. Кожен із 17-ти прийомів аналізується окремо, що дозволяє отримати числові оцінки їхньої вираженості у текстах. Відповідно ідея запропонованого методу полягає у використанні архітектур глибокого навчання, зокрема моделей на основі BERT, доповнених додатковими числовими векторами, які виражають силу прояву семантичних маркерів (рис. 3.9). Основна ідея полягає в тому, щоб покращити точність класифікації текстів, що містять пропагандистські прийоми, шляхом поєднання контекстуальної інформації з тексту та специфічних ознак (маркерів), які вказують на різні емоційні або маніпулятивні елементи. Результат у вигляді нейромережевої оцінки наявних прийомів пропаганди буде мати часткову додаткову поясненість у вигляді нейромережевих оцінок наявності семантичних маркерів.

Отже, відбувається формування навчального набору даних для виявлення нейромережами прийомів пропаганди шляхом перетворення вхідних даних у вигляді множини навчених нейромереж до оцінки сили прояву кожного маркера з доповненої множини маркерів та множини навчальних текстів для кожного прийому пропаганди у вихідні дані у вигляді розміченої множини навчальних текстів для кожного прийому пропаганди за силою прояву кожного з маркерів із доповненої множини маркерів.

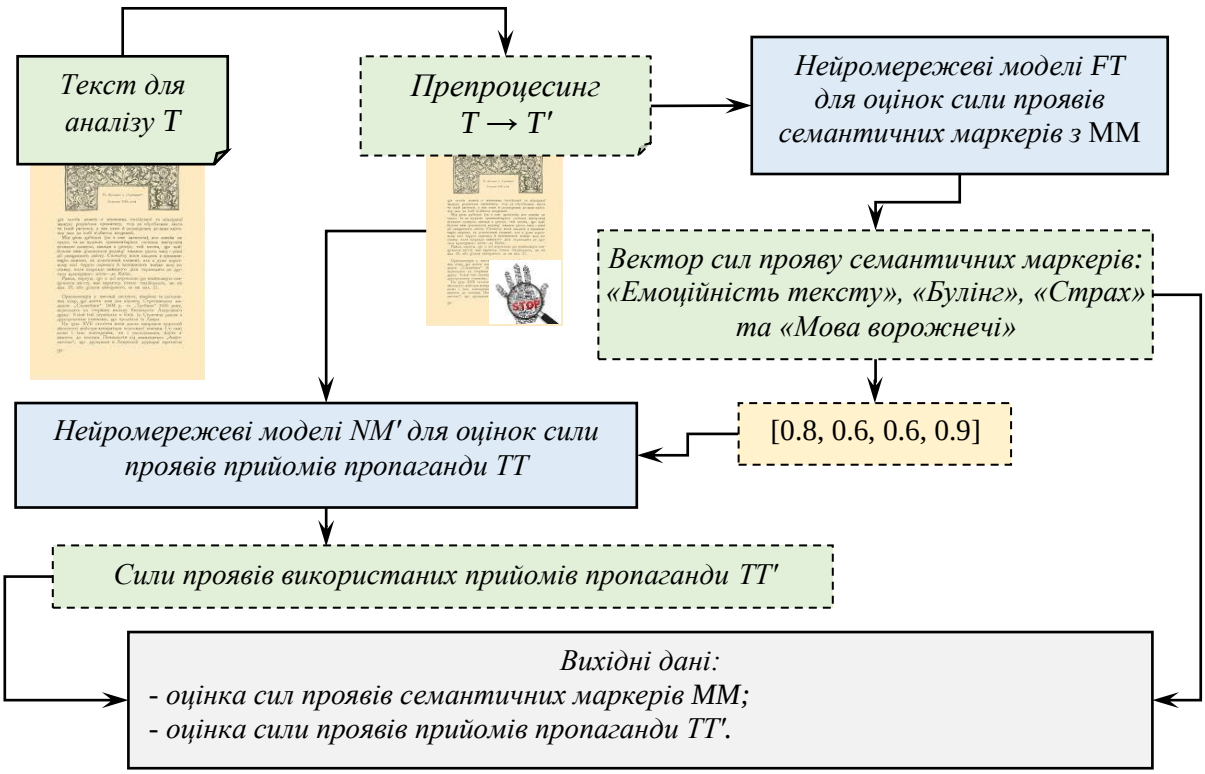


Рис 3.9 – Ілюстрація процесу перетворення даних для виявлення прийомів пропаганди за маркерами

Архітектура моделі-трансформера включає два основні компоненти: модель BERT, яка використовується для отримання контекстуальних векторних представлень тексту, і додатковий шар для обробки числових семантичних ознак. Модель BERT кодує текст, враховуючи як попередні, так і наступні слова, що дозволяє глибше зрозуміти семантичні зв’язки. Одночасно числові семантичні маркери проходять через окремий шар, який виділяє важливі ознаки та зменшує їх розмірність.

Після цього виходи з обох компонентів об'єднуються та передаються до кінцевого класифікатора, який визначає ймовірність наявності пропагандистських прийомів у тексті. Такий підхід дозволяє моделі враховувати як глибинні семантичні зв'язки, так і специфічні емоційні та маніпулятивні маркери, що значно покращує точність і надійність виявлення прийомів пропаганди.

Такими чином, розроблений метод виявлення прийомів пропаганди дозволяє не лише оцінити наявність кожного прийому пропаганди у тексті, а також отримати візуальну поясненість отриманих результатів.

3.3. Метод виявлення об’єктів пропаганди з візуальною інтерпретацією прийнятих рішень

Відповідно до п. 2.1 використані в тексті прийоми пропаганди можуть бути детектовані одним із існуючих нейромережевих підходів, при цьому вторинним результатом будуть відповідні нейромережеві моделі, навчені детектуванню окремих прийомів пропаганди. Запропонований підхід до виявлення об’єктів пропаганди з їх співставленням до використаних прийомів наведено на рис 3.10.

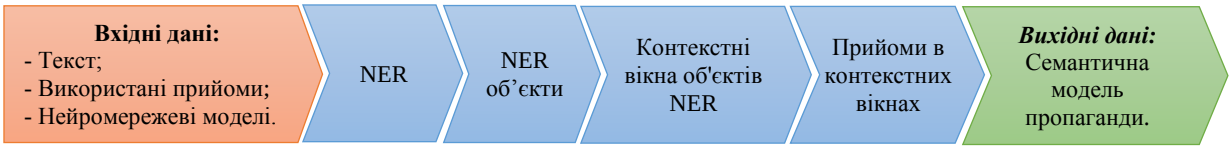


Рис. 3.10 – Підхід до ідентифікації об’єктів пропаганди

Цей підхід дозволяє забезпечити комплексний аналіз взаємозв’язків прийомів та об’єктів пропаганди в текстах, а також узагальнення для об’єктів пропаганди та їх альтернативних згадувань в текстах [136]. Відрізняється від існуючих аналогів тим, що об’єднує дві альтернативні задачі: пошук об’єктів пропаганди та співвіднесення знайдених об’єктів із числовими оцінками до знайдених прийомів.

У межах запропонованого підходу використано 17 окремих попередньо навчених нейромережевих моделей архітектури трансформер, які дозволяють визначати 17 вищеописаних основних прийомів пропаганди.

Запропоновані нейромережі навчені на розмічених даних, зібраних командою «Analysis Project» [137], яка провела аналіз текстів, виявивши всі фрагменти, які містять пропагандистські прийоми, а також їх тип. Зокрема, командою створено корпус новинних статей, анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів. Набір даних налічує 788 статей.

Створений відповідно до запропонованого підходу метод детектування об’єктів пропаганди [138] використовує нейромережеві засоби та дозволяє виявляти приналежність об’єктів до застосованих прийомів пропаганди, використовуючи подання семантичної моделі пропаганди *SMP* у тестовому тексті *T* (2.1)-(2.3)

Метод виявлення об’єктів пропаганди використовує нейромережеві моделі глибокого навчання та містить етапи: побудова набору об’єктів пропаганди *PO* шляхом розпізнавання іменованих сутностей; попередня обробка тексту та розширення *PO* за рахунок слів-представлень об’єктів; побудова контекстних вікон *CW* і їх об’єднання за *PO′*; виявлення рівня використання пропагандистських прийомів у *CW′* нейромережевими моделями *NM′*; побудова множини важливості віднесення *RTO′* між *TT′* та об’єктами *PO′* для тестового тексту *T* (рис. 3.11).

Метод виявлення об’єктів пропаганди вхідними даними має тестовий текст для виявлення об'єктів пропаганди – *T*; множину використаних прийомів у тестовому тексті – *TT′* і множину навчених нейромережевих моделей *NM′* для кожного прийому пропаганди.

Першим етапом є пошук іменованих сутностей (*NER*). Оскільки іменовані сутності можуть містити повтори, на цьому етапі всі повтори видаляються на рівні лем. Вихідними даними першого етапу є множина об’єктів пропаганди *PO* за *NER* без повторів.

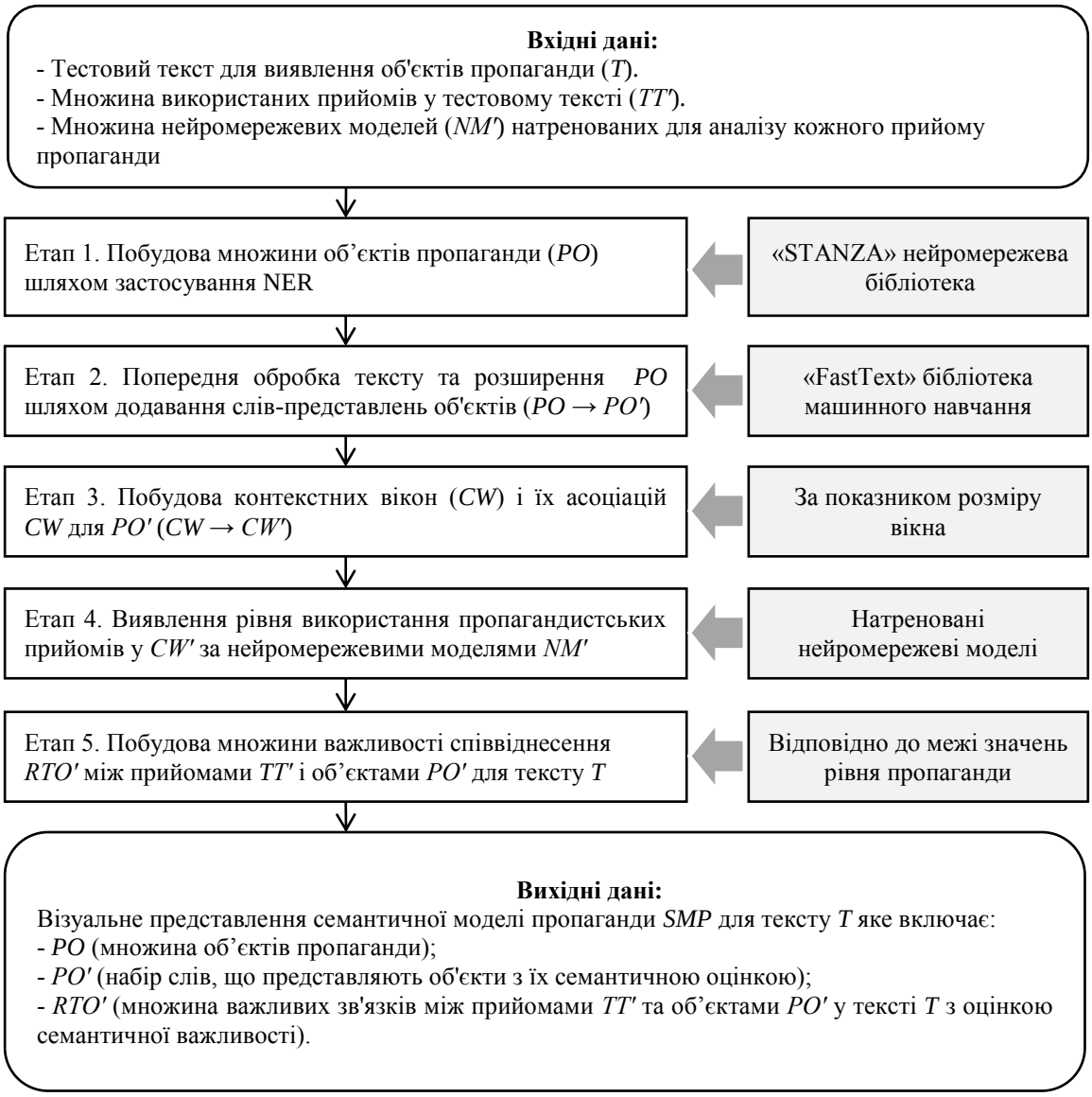


Рис. 3.11 – Схема методу виявлення об’єктів пропаганди

Другим етапом до кожної іменованої сутності є здійснення пошуку близьких за значеннями слів-об’єктів. Така потреба виникає тому, що поняття «об’єкти пропаганди» є дещо ширшим поняттям, ніж NER, що містяться в множині *PO* та включають також аспекти культури, групи об’єктів, об’єднаних за певними ознаками тощо.

Для пошуку схожих до NER об’єктів буде застосовано попередньо навчена модель «FastText» [139], яка розроблена Facebook AI Research.

«FastText» підтримує моделі «CBOW» і «Skip-gram», що дає змогу ефективно аналізувати контекст слів і виявляти семантичні зв'язки між ними.

Використання «FastText» у цьому контексті є доречним, оскільки модель дозволяє виявляти схожі слова та об'єкти на основі контекстуальних векторів, що є корисним для розширення спектра виявлених об'єктів пропаганди за межами іменованих сутностей. Модель «FastText» донавчається перед використанням на пропагандистських текстах. Вихідними даними етапу є розширення *PO* за рахунок семантично-близьких до об’єктів слів *PO′*. Мінімальна семантична близькість задається в залежності від задачі та визначається емпірично. У даному дослідженні поріг не застосовувався.

Наступним етапом є побуда контекстних вікон *CW* для кожного об’єкта пропаганди з *PO′*. Під контекстним вікном у межах роботи слід розуміти речення, де зустрічається вказаний об’єкт пропаганди. Якщо одне контекстне вікно містить декілька об’єктів пропаганди – вікна не дублюються (для 1 об’єкта пропаганди). По *CW′* виконується виявлення рівня використання пропагандистських прийомів у *CW′* нейромережевими моделями *NM′*. Оцінка приналежності контекстних вікон до використаних прийомів виконується шляхом векторизації контекстних вікон *CW′* відповідними векторизаторами *NM′* і аналізується його приналежність до кожного із проявлених у тексті прийомів.

Останнім етапом є побудова множини відношень *RTO′* між прийомами пропаганди *TT′* та об’єктами *PO′* для тексту *T*. Якщо сила прояву прийому пропаганди нижча певного порогового значення в межах оцінки елемента множини *CW′*, прийом не вважається застосованим до групи об’єктів.

Вихідними даними запропонованого підходу є візуальне представлення [140, 141] семантичної моделі пропаганди *SMP* для тексту *T* яке містить: множину об’єктів пропаганди; множину слів, що представляють об'єкти з їх семантичною оцінкою; множину важливості зв’язків між прийомами *TT′* та об’єктами пропаганди *PO′* у тексті *T* з оцінками семантичної важливості.

Отже, розроблений підхід дозволяє виявляти об’єкти пропаганди, які виходять за межі виявлення NER, а також дозволяє оцінити приналежність виявлених об’єктів до застосованих у тексті пропагандистських прийомів.

3.4. Джерела даних дослідження

Для навчання моделей рекурентних нейронних мереж було сформовано набір даних з понад 25000 дописів, які були розмічені відповідно до приналежності до категорій «Пропаганда» та «Не пропаганда». Переліки пропагандистських та верифікованих джерел було сформовано згідно з офіційними каналами Президента й Верховної Ради України, а також з даними аналітичних міжнародних авторитетних досліджень [142, 143] та аналітичних зведень [144].

Для нормалізації вхідних даних було відкинуто записи довжиною менше 200 і більше 6300 символів. У результаті фільтрації даних отримано набір, що складається із 21222 елементів, де 10737 записів належать класу «пропагандистський допис» та 10485 записів – класу «допис без пропаганди».

Для нормалізації вхідних даних було відкинуто записи довжиною менше 200 символів і більше 6300 символів. У результаті фільтрації даних отримано набір, що складається із 21222 елементів, де 10737 записів належать класу «пропагандистський допис» та 10485 записів – класу «допис без пропаганди». Графік розподілу даних за довжиною в символах наведено на рис 3.12.

Відповідно до рис. 3.12 більшу кількість записів мають тексти, які не містять пропаганди й знаходяться в діапазоні довжини 200..800 символів, що може негативно вплинути на якість класифікації у майбутньому.

Водночас набір пропагандистських текстів є більш рівномірно розподіленим. Усі багатомовні фрагменти було автоматично перекладено на українську мову.

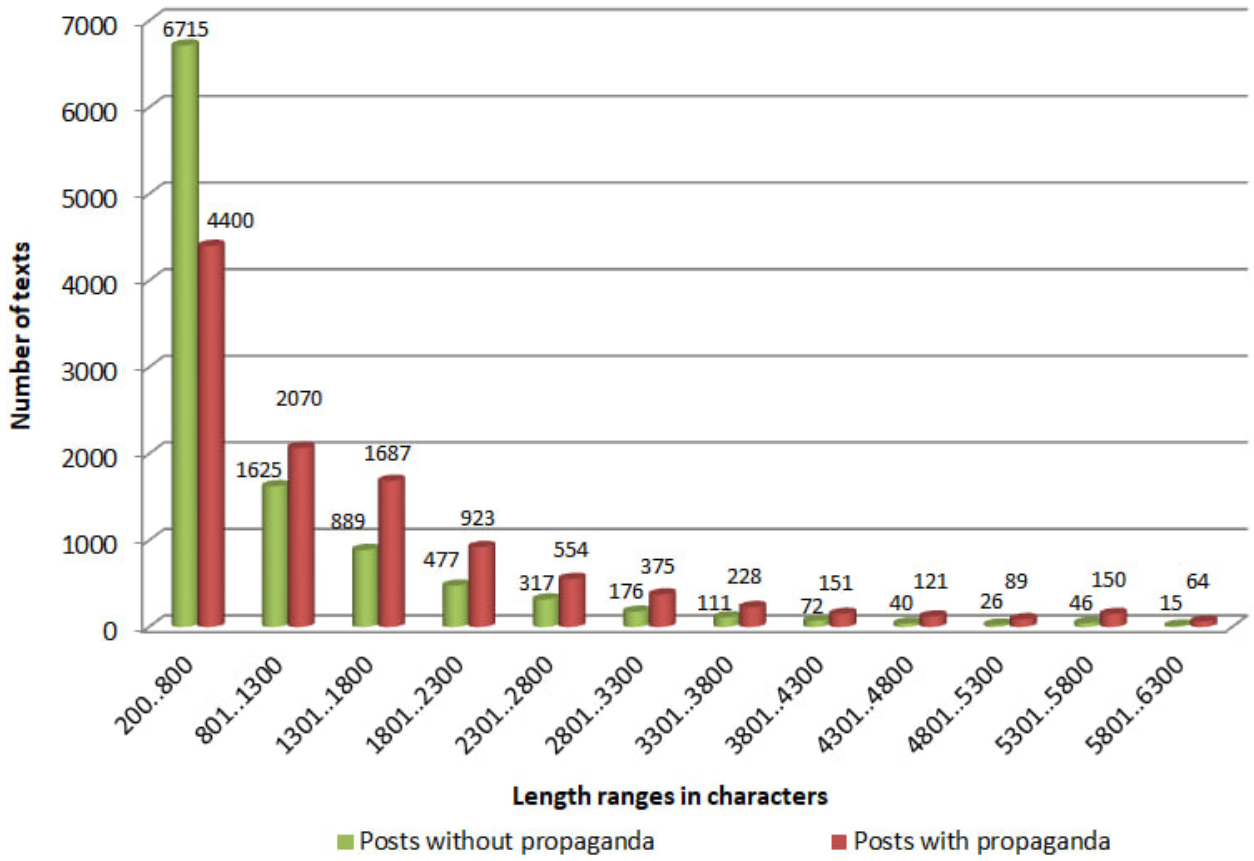


Рис. 3.12 – Розподіл даних за кількістю символів у об’єднаній вибірці

Описаний набір даних буде використано для навчання моделей нейронних мереж у межах розробленого методу виявлення пропаганди.

Для навчання моделей машинного навчання, що будуть виконувати функції виявлення прийомів пропаганди, використано набір даних «emnlp_trans_uk_dataset», що є перекладеним набором даних «emnlp_en_dataset» з відповідністю розмітки українською мовою, взятий з Kaggle-змагань «Disinformation Detection Challenge» [145] з посиланням на «Analysis Project» та Zenodo [145]. Команда «Analysis Project» [137] провела аналіз текстів, виявивши всі фрагменти, які містять пропагандистські прийоми, а також їх тип. Зокрема, ними створено корпус новинних статей, анотованих вручну на рівні фрагментів

за допомогою вісімнадцяти пропагандистських прийомів. Набір даних налічує 788 статей. Розподіл статей за довжиною у символах наведено на рис. 3.13.

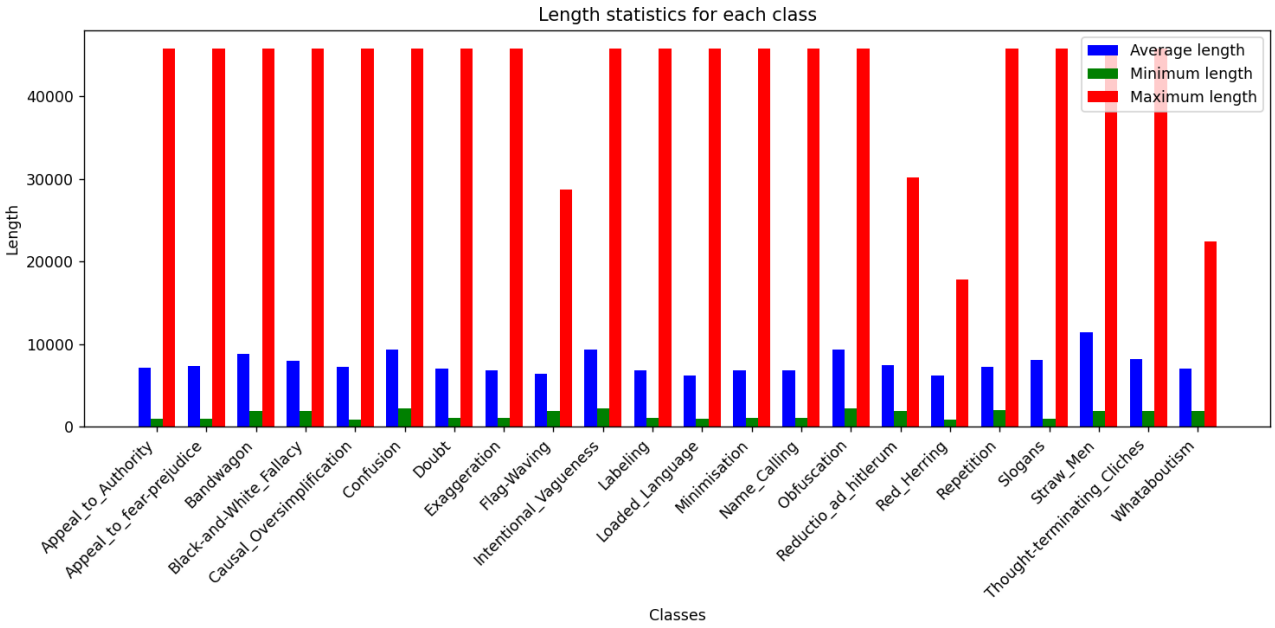


Рис. 3.13 – Статистика за довжиною у символах за прийомами пропаганди у [144]

Відповідно до графіка на рисунку 3.13, для більшості прийомів пропаганди довжина текстів, де вони представлені особливої ролі не грають. Однак «Розмахування прапором», «Червоний оселедець», «Зведення до Гітлера» та «Ватабоутізм» все ж мають меншу максимальну довжину в текстах, де вони представлені.

Для тренування моделей машинного навчання даний датасет було модифіковано таким чином, щоб текст, який містить кожен прийом пропаганди, був розміщений в окремому каталозі. Після такого перерозподілу виведено статистику наявних текстів, що репрезентують прийоми пропаганди. Статистика наведена на рис. 3.14.

Відповідно до рис. 3.14 деякі прийоми пропаганди, такі як «Bandwagon», «Confusion», «Intentional Vagueness», «Obfuscation» та «Straw Men», представлені у критично низькій кількості (менше 20 тестів), тому для них

окремі класифікатори не створювались, ці дані об’єднані у категорію «Інші прийоми пропаганди», однак таким чином, щоб у наявному наборі не були присутні інші прийоми, відмінні від п’яти перерахованих. До прийомів пропаганди, що представлені менш ніж у 100 документах, але більше ніж у 20-ти буде застосовано SMOTE-балансування під час навчання класифікаторів. До таких категорій належать: «Апеляція до авторитету», «Помилка хибної дихотомії», «Зведення до Гітлера», «Червоний оселедець», «Гасла», «Кліше, що припиняють думання» та «Ватабоутізм».

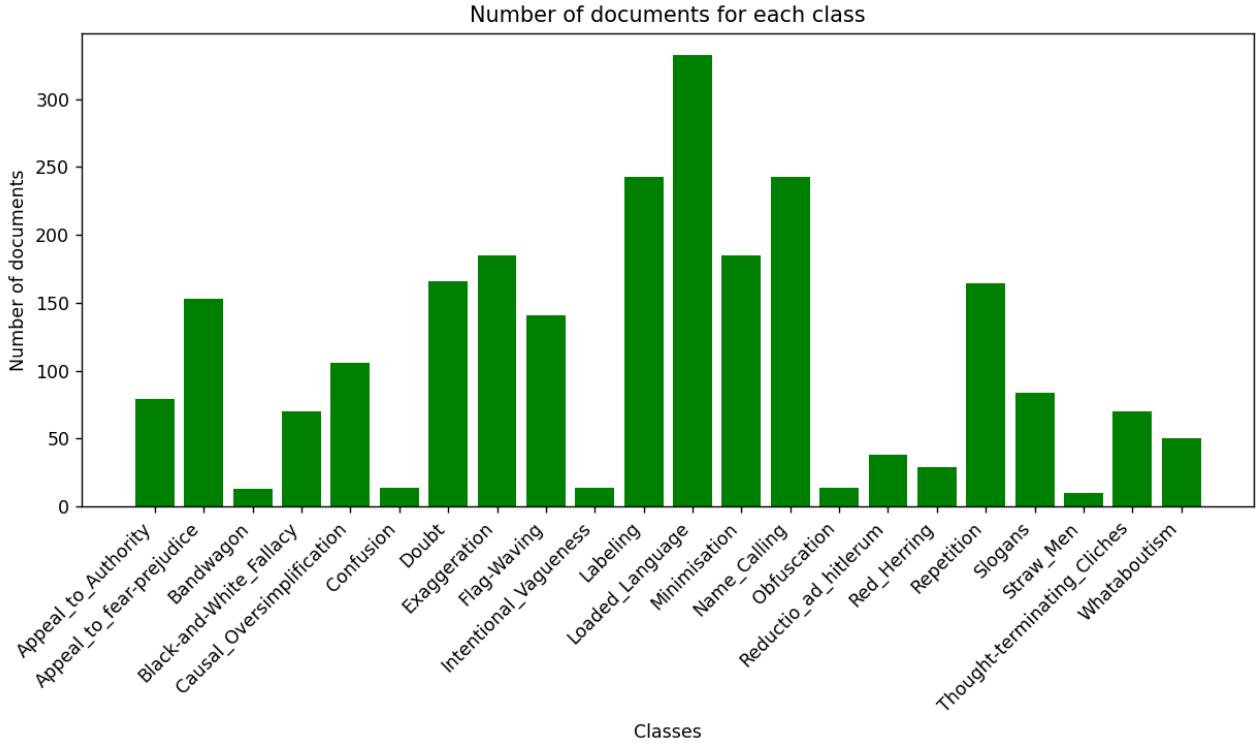


Рис. 3.14 – Статистика за кількістю текстів, що представляють прийоми пропаганди у [145] (в шт.)

В якості зразків для аналізу було використано дописи із соціальних мереж, які були пропрацьовано «Центром стратегічних комунікацій» [147]. Такі дописи є вже розміченими та містять висновки експертів, які можна порівняти з вихідними даними роботи запропонованого підходу.

3.5. Висновки до розділу 3

У третьому розділі дисертаційної роботи було розроблено та представлено методи виявлення прийомів і об’єктів пропаганди з використанням нейромережевих технологій. Ці методи спрямовані на оцінку текстового контенту, визначення сили проявів пропагандистських прийомів та забезпечення візуальної інтерпретації отриманих результатів.

Основною метою методу виявлення прийомів пропаганди за маркерами є оцінка текстового контенту на предмет наявності пропагандистських прийомів та визначення сили їх проявів. Відмінність запропонованого методу від існуючих полягає у використанні додаткової множини маркерів при навчанні нейромережевих класифікаторів, що дозволяє здійснювати візуальну інтерпретацію результатів. Маркери містять такі ознаки, як «Емоційність тексту», «Булінг», «Страх» та «Мова ворожнечі», що дозволяє точно ідентифікувати конкретні пропагандистські прийоми.

Метод виявлення об'єктів пропаганди з візуальною інтерпретацією дозволяє забезпечити комплексний аналіз взаємозв'язків прийомів та об'єктів пропаганди в текстах, а також узагальнення для об'єктів пропаганди та їх альтернативних згадувань. Використання 17-ти попередньо навчених нейромережевих моделей архітектури трансформер дозволяє точно визначати прийоми пропаганди та виявляти об’єкти, які виходять за межі виявлення NER.

Розроблені методи виявлення прийомів і об’єктів пропаганди забезпечують інструмент для аналізу текстового контенту і дозволяють не лише ідентифікувати наявність пропаганди, а й забезпечити візуальну інтерпретацію отриманих результатів, що є засобом забезпечення інформаційної безпеки та протидії маніпулятивному впливу.

РОЗДІЛ 4.
ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ ПРИЙОМІВ ТА ОБ'ЄКТІВ ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ

4.1. Опис експериментального програмного забезпечення

Для *дослідження методу класифікації текстів за вмістом пропаганди з використанням нейромережевих RNN-моделей BiLSTM та GRU* створено програмну реалізацію засобами мови Python. Інтерфейс програмної складової, що відповідає за модуль навчання нейромережевих RNN-моделей BiLSTM та GRU, наведено на рисунку 4.1 (а). Інтерфейс програмної частини, що відповідає за процес виявлення пропаганди розробленим методом, наведено на рисунку 4.1 (б).

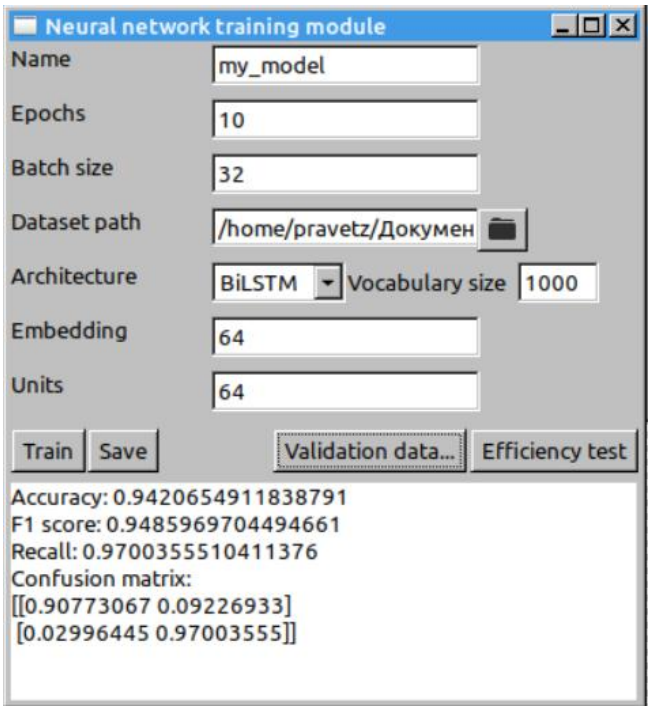


Рис. 4.1 (а) – Модуль навчання RNN-моделей BiLSTM та GRU

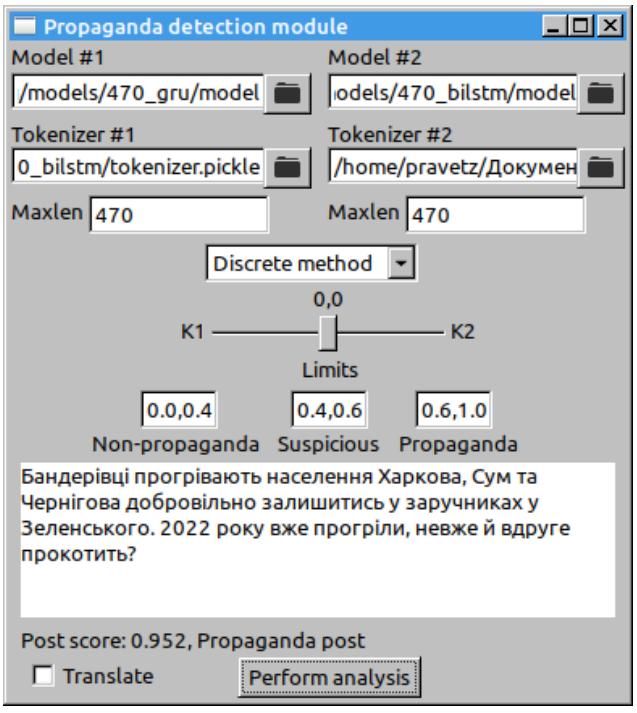


Рис. 4.1 (б) – Модуль виявлення пропаганди

Програмна реалізація дозволяє здійснювати навчання моделей нейронних мереж та використовувати їх для виявлення пропаганди в текстовому інтернет-контенті.

Для оцінки ефективності розробленого *методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання (на основі нейромережі гібридної архітектури)* створено програмну реалізацію, яка складається із ноутбуків, реалізованих у хмарному сервісі «Google Colab» (для навчання нейромережевої моделі BiLSTM гібридної архітектури та для розширення отриманого набору даних методом аугментації тексту), а також застосунок з графічним інтерфейсом користувача на мові Python з використанням середовища розробки PyCharm. Приклад роботи розробленого застосунку наведено на рис. 4.2.

Для *дослідження методу виявлення прийомів пропаганди за маркерами* з візуальною інтерпретацією прийнятих рішень та *методу виявлення об'єктів пропаганди* з візуальною інтерпретацією прийнятих рішень створено спільну програмну реалізацію, оскільки результатом є не лише виявлені прийоми та об'єкти пропаганди, а й їх взаємозв'язок.

Експериментальне програмне забезпечення для виявлення прийомів та об'єктів пропаганди складається із наборів ноутбуків, реалізованих у хмарному сервісі «Google Colab», які призначені *для навчання нейромережевих моделей* BERT із подальшим збереженням їх на жорсткому диску для використання у вебзастосунку для виявлення прийомів пропаганди, а також набору ноутбуків для збереження нейромережевих моделей для виявлення сили прояву маркерів пропаганди.

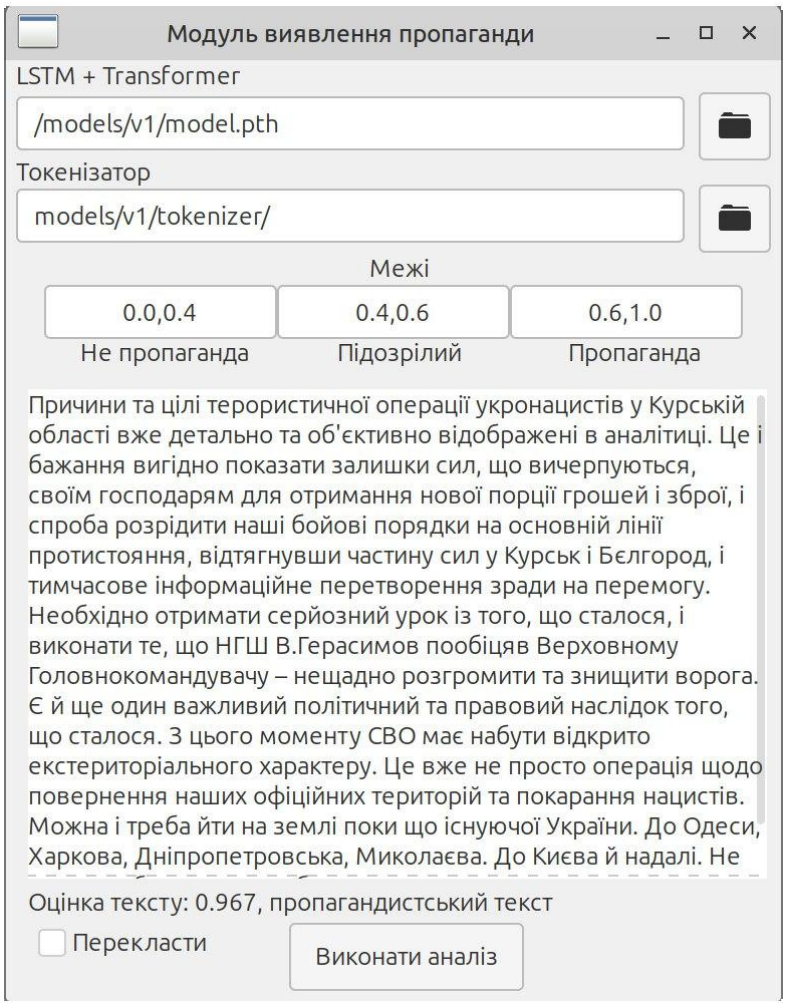


Рис. 4.2 – Програмна реалізація методу класифікації текстів за вмістом пропаганди з використанням нейромережі BiLSTM гібридної архітектури

Для *виявлення прийомів та об'єктів пропаганди* навченими нейромережевими моделями було створено експериментальне програмне забезпечення у вигляді вебзастосунку засобами мови програмування Python за допомогою середовища розробки PyCharm. Вебзастосунок використовує 17 попередньо навчених нейромережевих моделей-трансформерів, нейромережеву бібліотеку «Stanza» для пошуку NER, фреймворк «Flask», попередньо навчену модель «FastText», яка проходить донавчання на пропагандистських текстах. Інтерфейс експериментальної установки наведено на рисунку 4.3.

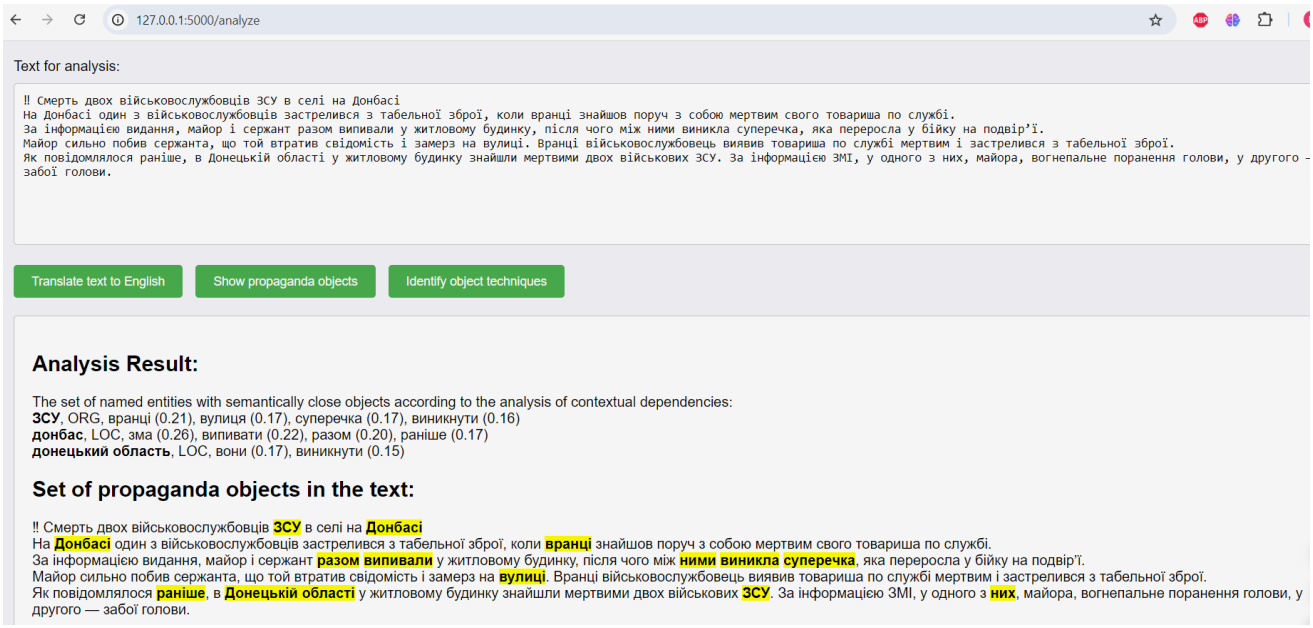


Рис. 4.3 – Приклад отримання результату за допомогою візуальної аналітики розробленим експериментальним програмним забезпеченням

Створене експериментальне програмне забезпечення має функціональність визначати використані прийоми за текстом, визначати об'єкти пропаганди та аналізувати приналежності виявлених об'єктів до використаних прийомів пропаганди.

4.2. Дослідження методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання

Дослідження методу класифікації текстів за вмістом пропаганди з нейромережевими моделями BiLSTM та GRU [148] проводилося на основі розробленого експериментального програмного забезпечення [149].

Під час дослідів нейромережі навчались із різними параметрами (batch, epoch), результати порівняння найкращих моделей глибокого навчання наведені у таблиці 4.1.

Таблиця 4.1

Залежність метрик від параметрів нейромереж

Параметри:	**GRU**		**BiLSTM**	
batch	32	64	32	64
epoch	20	20	20	20
Метрики:				
Точність (Accuracy)	0.97	0.96	0.96	0.95
Втрати	0.04	0.06	0.04	0.07

Відповідно до таблиці 4.1 GRU має вищу точність, ніж BiLSTM при однакових параметрах. На Рис. 4.4 (а) наведено розподіл коректно класифікованих текстів нейромережею GRU, а на Рис. 4.4 (б) – розподіл некоректно класифікованих текстів.

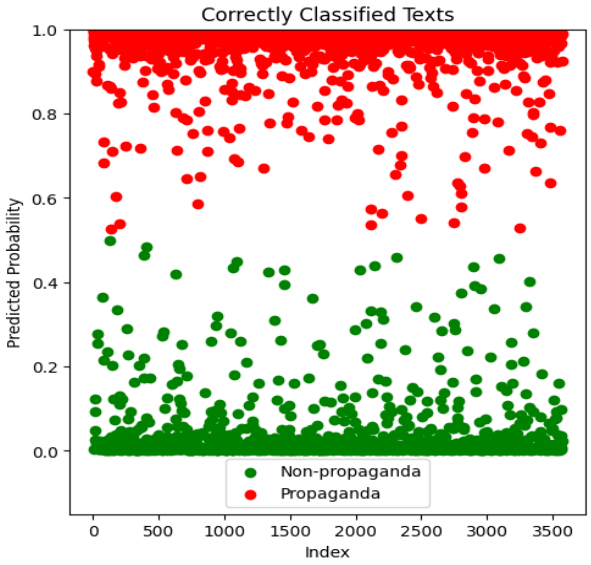


Рис. 4.4 (а) – Розподіл коректно класифікованих текстів нейромережею GRU

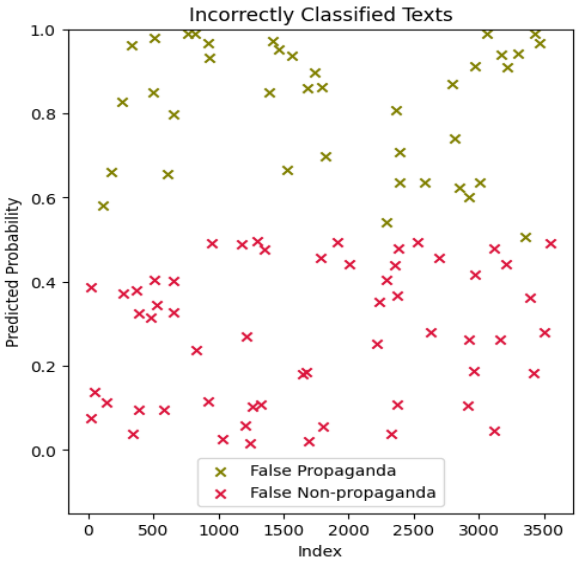


Рис. 4.4 (б) – Розподіл некоректно класифікованих текстів нейромережею GRU

В якості валідаційних даних було використано 3573 записи, з яких до класу «пропагандистський текст» належав 1951 запис, а до класу «без

пропаганди» – 1622 записи. З них коректно було класифіковано 1912 записів класу «пропагандистський» та 1565 записів класу «без пропаганди».

Однак 57 текстів класу «без пропаганди» нейромережею хибно класифіковано як пропаганда, а 39 текстів класу «пропагандистський» хибно класифіковані як «без пропаганди». Загальна точність на валідаційних даних становить 0.97. Відповідно до числових даних, клас «без пропаганди» класифікується дещо гірше, ніж клас «пропагандистський».

На Рис. 4.5 (а) наведено розподіл коректно класифікованих текстів нейромережею BiLSTM, а на Рис. 4.5 (б) – розподіл некоректно класифікованих текстів.

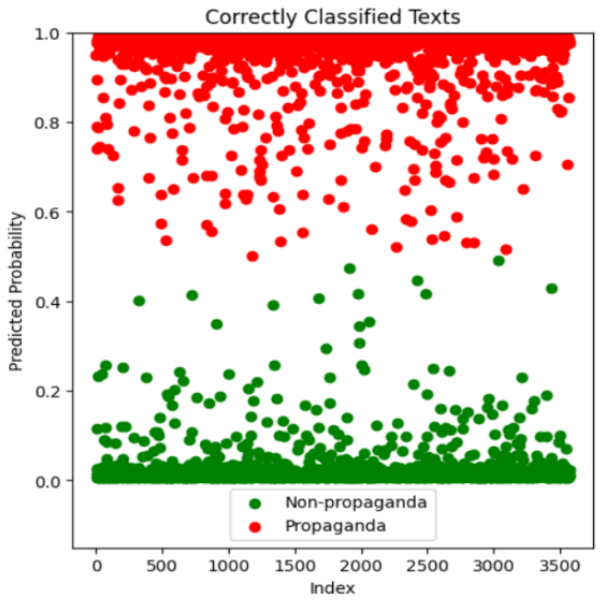


Рис. 4.5 (а) – Розподіл коректно класифікованих текстів нейромережею BiLSTM

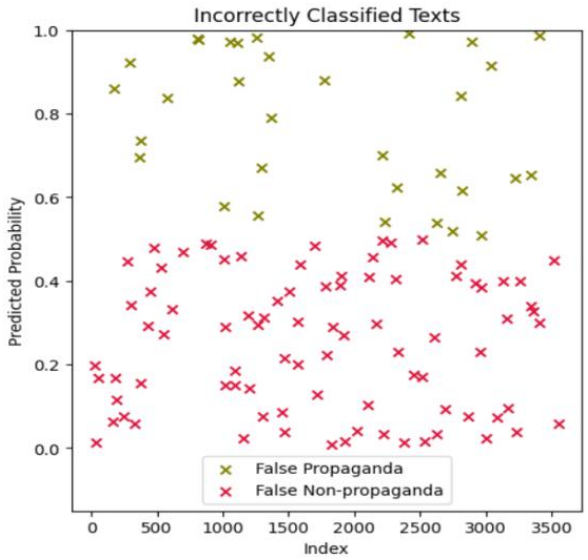


Рис. 4.5 (б) – Розподіл некоректно класифікованих текстів нейромережею BiLSTM

З 3573-х валідаційних дописів, коректно класифіковано 1883 дописи класу «пропагандистський» та 1572 тексти класу «без пропаганди»; 86 текстів класу «без пропаганди» нейромережею хибно класифіковано як пропаганда, а 32

тексти класу «пропагандистський» хибно класифіковані як «без пропаганди». Загальна точність на валідаційних даних становить 0.967.

Відповідно до Рис. 4.4 (а) та Рис.4.5 (а), тексти мають достатньо високий рівень міжкласової роздільності, водночас відповідно до Рис. 4.4 (б) та Рис. 4.5 (б) некоректно класифіковані дані зосереджені ближче до центральної частини графіків, що свідчить про доцільність підходу з розбиттям на 3 класи: «пропагандистський», «без пропаганди», «підозрілий».

Для дослідження ефективності виявлення пропаганди в текстовому інтернет-контенті розробленим методом було використано метрики Acuracy, Precision, Recall та F1, описані в п. 2.6. Значення метрик для дискретного та бінарного варіацій методу наведено в Таблиці 4.2. Хоча бінарний підхід дав гірші результати за метрикою Acuracy, однак дав кращі показники за метриками Precision, Recall та F1. Водночас дискретний підхід Acuracy практично не погіршив, але при цьому метрики Precision, Recall та F1 дещо йому поступаються.

Таблиця 4.2

Значення метрик за бінарним та дискретним підходами

Підхід	**Acuracy**	**Precision**	**Recall**	**F1**
Дискретний	0.97	0.973	0.981	0.976
Бінарний	0.95	0.977	0.987	0.981

Для експерименту задано такі параметри дискретного підходу: k_1=0.5, k_2=0.5, l_2 = 0.45, l_4 = 0.55.

Хоча бінарний підхід дав гірші результати за метрикою Acuracy, однак дав кращі показники метрик Precision, Recall та F1, водночас дискретний підхід Acuracy практично не погіршив, але при цьому метрики Precision, Recall та F1 дещо йому поступаються. Діаграма значень метрик наведена на Рис. 4.6.

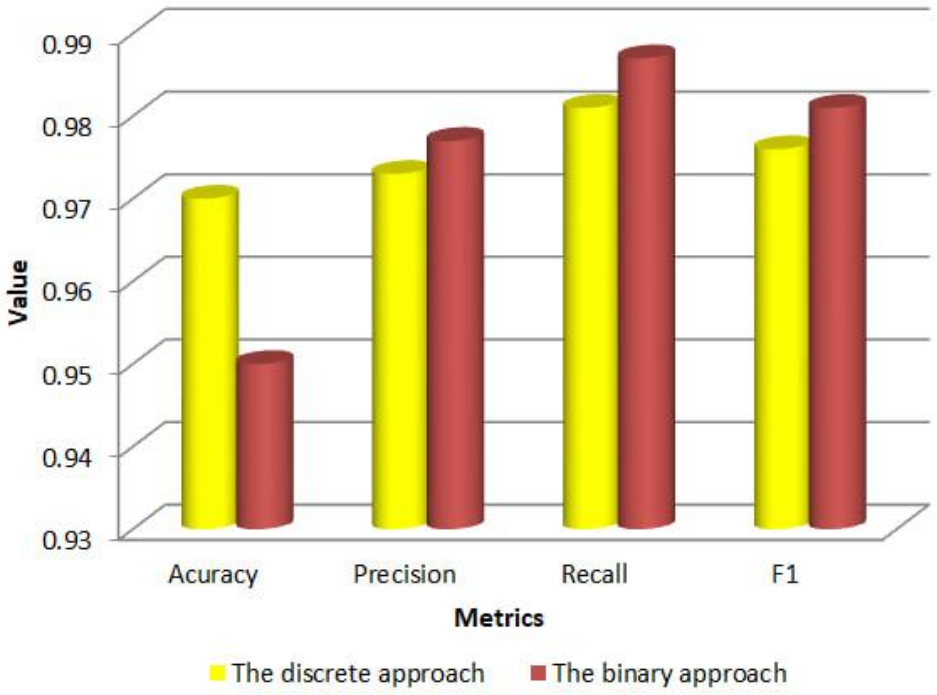


Рис. 4.6 – Значення метрик для бінарного та дискретного підходів до класифікації текстів за вмістом пропаганди

Для експерименту задані такі параметри дискретного підходу: k_1=0.5, k_2=0.5, l_2 = 0.45, l_4 = 0.55.

Однак перевагами дискретного методу є його гнучкість і можливість налаштовування залежно від задачі. Тому експерименти у даному напрямі є перспективними.

Щодо текстів, які були ідентифіковані як підозрілі, то в них виявлено специфічні ознаки. Наприклад: «*хай л-т спільнота успокояться . Вони ж начеб то такі самі люди (туди ж і націоналістів) , а значить над ними можна жартувати як і над усіма іншими 2) Менеджери то сатира тому агонь .3) Україн на щастя не США , тому можна жартувати і над реальними випускниками тернопольської медки.П. С. Найкращий жарт де мусоріна зупиняє Беста , бере хабар . А той потім повертається і знімає його на гарячому.П. С. С. Де огляд скетчу з Хмельницьким та послом московським ? Там один фак американський чтого тільки вартий !*». Текст містить ряд слів-

тригерів, як-от «націоналістів», «московським», а контекст є схожим до пропагандистського.

Також у наборі даних зустрічались помилково віднесені тексти. Наприклад, наступний текст у датасеті був помічений як «допис без пропаганди», однак його вміст «*Росія не претендує на територію України, якби Україна не напала на Росію, жодних військових дій не було б. Вірніше на Росію напали країни НАТО на території України, тому що влада України продала країну і зрадила свій народ і погодилася воювати за інтереси Байдена, Макрона, Сунака та Шольца до останнього живого українця.*» є відвертою пропагандою, й обидві нейромережеві моделі дуже високо оцінили наявність у ньому пропаганди, давши оцінки 0.92 (GRU) та 0.97 (BiLSTM); загальна оцінка – 0.944.

Запропоновано метод виявлення пропаганди в текстовому інтернет-контенті нейромережевими засобами обробки природної мови, який працює з українською мовою, а також проведена його апробація.

У межах дослідження виконано тестове навчання ансамблю класифікаторів із нейромережевих архітектур BiLSTM та GRU; розроблено програмне забезпечення, що імплементує створений метод, та проведено дослідження його ефективності.

Запропонований підхід здатний визначати пропаганду ансамблем моделей RNN з показниками Acuracy 0.97, Precision 0.973, Recall 0.981 і F1 0.976 при дискретному підході та Acuracy 0.95, Precision 0.977, Recall 0.987 та F1 0.981 при бінарному підході.

Розроблений метод має обмеження: працює з текстовими дописами довжиною від 200 до 6300 символів. Для коротших та довших текстів спостерігається погіршення продуктивності.

Дослідження методу класифікації текстів за вмістом пропаганди з нейромережевою моделлю BiLSTM гібридної архітектури проведено із

застосуванням розширення отриманого набору даних методом аугментації та без розширення набору даних. У ході дослідження вдалось досягнути точності (Acuracy) 0.98. Результати експерименту наведені в таблиці 4.3 та 4.4.

При використанні лише 2-х категорій – «пропаганда» та «не пропаганда» – найбільша кількість помилок зосереджена в проміжку 0.4 – 0.6 (рис. 4.7). Тому уведення категорії «підозрілий текст» із можливістю плаваючих меж, що адаптивні предметній області є доцільним.

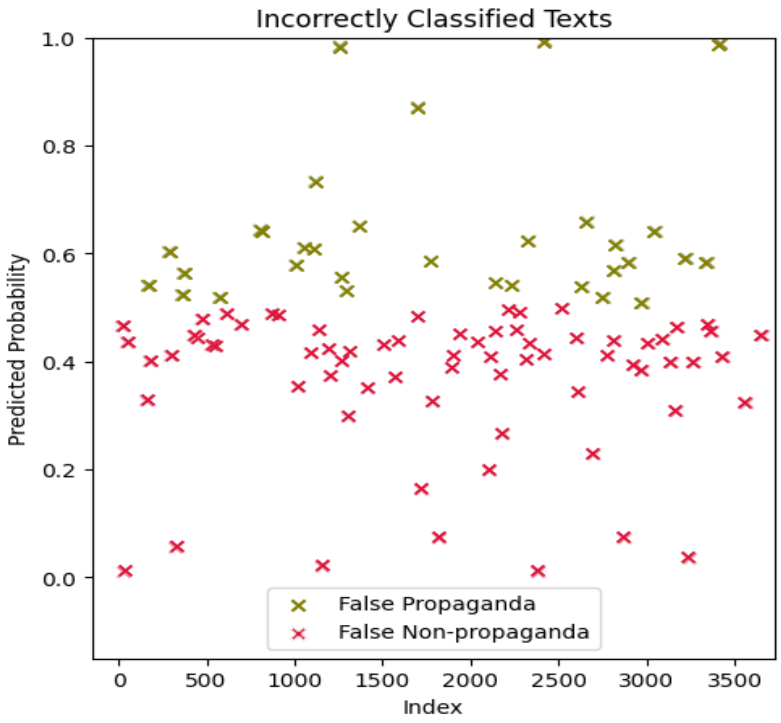


Рис. 4.7 – Розподіл некоректно ідентифікованих зразків текстів до категорій «пропаганда» та «не пропаганда»

Завдяки уведенню категорії «підозрілий текст» показник Accuracy незначно зменшився, однак показники Precision та Recall зросли, що свідчить про можливість ефективної автоматизованої модерації текстів на предмет пропаганди.

Таблиця 4.3

Результати навчання нейромережі BiLSTM гібридної архітектури без використання аугментації навчальної вибірки

Кількість епох:	**Accuracy**	**Precision**	**Recall**
5	0.923	0.935	0.948
6	0.934	0.942	0.953
7	0.931	0.94	0.952
8	0.952	0.955	0.965
9	0.962	0.96	0.97
10	0.962	0.961	0.971
15	**0.971**	**0.965**	**0.978**
20	0.965	0.963	0.976

Accuracy розраховано як відсоток правильних класифікацій серед усіх чітко класифікованих текстів (без урахування категорії «підозрілий»). Precision і Recall враховують тільки ті тексти, що були чітко ідентифіковані як «пропагандистський» та «без пропаганди».

Відповідно до таблиць 4.3 та 4.4 вдалося досягти Accuracy 0.971 при гібридній архітектурі без використання аугментації навчальної вибірки та 0.978 з використанням аугментації навчальної вибірки. Ілюстарція проведених експериментів за метрикою Accuracy наведена на рис. 4.8.

Відповідно до рис. 4.8 із застосуванням аугментації кращі показники досягаються при більшій кількості епох. Це пояснюється розширенням навчальної вибірки, що призводить до потреби більшої кількості епох. Водночас при застосуванні аугментації вдалося досягнути точності 0.978, при тому, що без аугменації цей показник максимально досяг рівня 0.971.

Таблиця 4.4

Результати навчання нейромережі BiLSTM гібридної архітектури з використанням аугментації навчальної вибірки

Кількість епох:	**Accuracy**	**Precision**	**Recall**
5	0.855	0.865	0.871
6	0.863	0.872	0.878
7	0.929	0.931	0.935
8	0.932	0.935	0.937
9	0.953	0.955	0.96
10	0.951	0.954	0.959
15	0.971	0.972	0.976
20	**0.978**	**0.982**	**0.982**

Отримані результати говорять про спроможність запропонованого методу ефективно класифікувати тексти за вмістом пропаганди нейромережевими моделями глибокого навчання. Застосування додаткової категорії «підозрілий текст» дозволило підняти показники Precision та Recall, що у свою чергу дає можливість автоматизованої модерації текстів на предмет пропаганди з помилками не більше 0.018 для хибного виявлення пропаганди.

Метод ґрунтується на об’єднанні традиційних рекурентних нейромереж з довгостроковою пам’яттю із трансформерами, що дозволяє забезпечити більш глибоке розуміння послідовності та контексту в текстовому контенті та дозволяє досягнути Accuracy 0.978.

У таблиці 4.5 наведено порівняння існуючих рішень щодо виявлення пропаганди у текстовому контенті з розробленим методом при різних варіантах нейромережевих архітектур глибокого навчання.

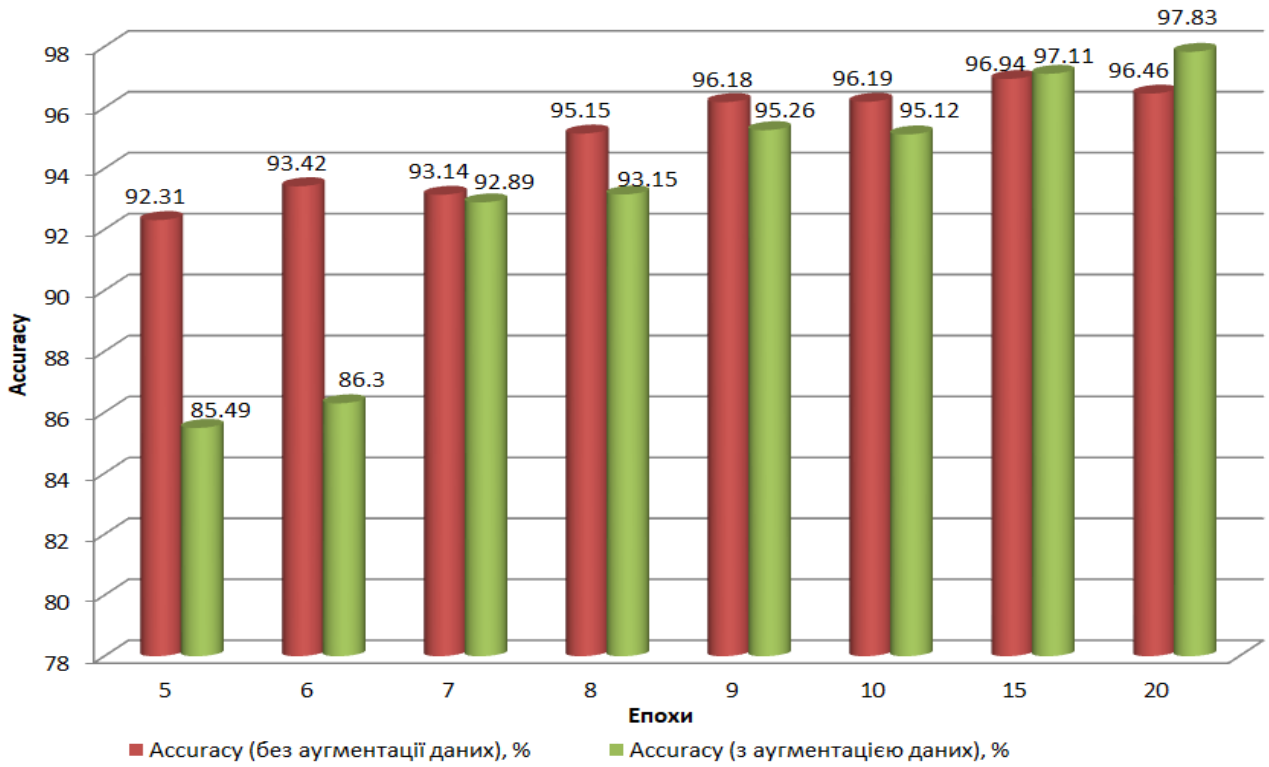


Рис. 4.8 – Порівняння показників Accuracy із аугментацією та без аугментації

Із проведеного дослідження видно, що із застосуванням аугментації кращі показники досягаються при більшій кількості епох, що пояснюється розширенням навчальної вибірки, яке призводить до потреби більшої кількості епох. Водночас при застосуванні аугментації вдалося досягнути точності 0.978; без аугментації цей показник максимально досяг рівня 0.971.

Порівнюючи отримані результати з відомими аналогами, вдалось досягнути підвищення точності на 0.007 при дискретній реалізації ансамблевого підходу, на 0.027 при бінарній реалізації ансамблевого підходу та на 0.035 при використанні нейромережі гібридної архітектури порівняно з методом на основі логістичної регресії, реалізованим в [45]. Як порівняти з метрикою F1 гібридної архітектури, вдалось досягнути підвищення на 0.06 порівняно з [35] (XLM RoBERTa Large зі стилометрією) та на 0.071 порівняно з [35] (XLM RoBERTa Large без стилометрії).

Таблиця 4.5

Порівняння існуючих підходів до виявлення пропаганди з розробленими

Назва методу, автор	**Accuracy**	**Precision**	**Recall**	**F1**
Логістична регресія (*Висоцька*) [45]	0.943	не вказано	не вказано	не вказано
GPT-4 base *(Kilian Sprenkamp, Daniel Gordon, Liudmila Zavolokina)* [94]	не вказано	0.529	0.645	не вказано
GPT-4 chain of thought *(Kilian Sprenkamp, Daniel Gordon, Liudmila Zavolokina)* [94]	не вказано	0.569	0.578	не вказано
XLM RoBERTa Large зі стилометрією (*Aleš Horák, Radoslav Sabol, Ondřej Herman, Vít Baisa*) [35]	не вказано	не вказано	не вказано	0.922
XLM RoBERTa Large без стилометрії (*Aleš Horák, Radoslav Sabol, Ondřej Herman, Vít Baisa*) [3534]	не вказано	не вказано	не вказано	0.911
Метод на основі ансамблів. Дискретний (*розроблений*)	0.95	0.977	0.987	0.976
Метод на основі ансамблів. Бінарний (*розроблений*)	0.97	0.973	0.981	0.981
Метод на основі нейромережі гібридної архітектури (*розроблений*)	0.978	0.982	0.982	0.982

Відповідно до таблиці 4.5 всі розроблені методи показують приріст за метриками, як порівняти з існуючими аналогами, однак підхід на основі нейромережі гібридної архітектури є кращим, оскільки показує вищі результати.

Одержані результати свідчать про спроможність запропонованого методу ефективно класифікувати тексти за вмістом пропаганди нейромережевими моделями глибокого навчання, а застосування додаткової категорії «підозрілий текст» дозволило підняти показники Precision та Recall, що у свою чергу дає

можливість автоматизованої модерації текстів на предмет пропаганди з помилками не більше 0.018 для хибного виявлення пропаганди.

4.3. Дослідження методу виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень

Дослідження методу виявлення прийомів пропаганди [150] за маркерами із візуальною інтерпретацією прийнятих рішень [151] у текстовому контенті засобами штучного інтелекту проводилось у розрізі трьох основних підходів: традиційного підходу на основі моделей машинного навчання, підходу на основі рекурентних нейронних мереж та підходу на основі нейромереж архітектури трансформер. Дослідження виконується на базі розробленого програмного забезпечення [152].

Перша частина дослідження спрямована на виявлення природних закономірностей пропагандистських прийомів без застосування числового вектора з силами проявів семантичних маркерів. Результати наведені у таблицях 4.6 – 4.11.

Результати дослідження традиційного підходу на основі моделей машинного навчання для виявлення прийомів пропаганди «Апеляція до авторитету», «Помилка хибної дихотомії», «Зведення до Гітлера», «Червоний оселедець», «Гасла», «Кліше, що припиняють думання» та «Ватабоутізм» без використання SMOTE для балансування за метрикою Accuracy наведено в таблиці 4.6.

Відповідно до таблиці 4.6 точність виявлення прийомів пропаганди коливається від 0.51 до 0.68 за метрикою Accuracy, що є доволі низьким показником.

Таблиця 4.6

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди до SMOTE для балансування за метрикою Accuracy

Прийоми пропаганди	**Regression**	**SVM**	**Random Forest**	**Naive Bayes**
Апеляція до авторитету	0.57	0.63	0.55	**0.64**
Помилка хибної дихотомії	0.55	**0.64**	0.51	0.58
Зведення до Гітлера	**0.68**	0.56	0.61	0.59
Червоний оселедець	**0.61**	**0.61**	0.59	**0.61**
Гасла	0.62	**0.63**	0.56	0.62
Кліше, що припиняють думання	0.59	0.58	**0.63**	0.58
Ватабоутізм	0.62	**0.65**	0.59	0.57

Наступним етапом для даних прийомів пропаганди застосовано SMOTE балансування та збільшення таким чином кількості навчальних зразків до рівня не менше 100 для кожного з прийомів пропаганди. Результат експерименту наведено у таблиці 4.7.

Відповідно до таблиці 4.7 застосування SMOTE для балансування дало позитивні результати для більшості прийомів пропаганди, однак для «Гасла» результат покращення не дав. Це пов’язано з тим, що кількість навчальних зразків близька до граничної і є достатньою для навчання запропонованих версій машинного навчання.

Таблиця 4.7

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди зі SMOTE для балансування за метрикою Accuracy

Прийоми пропаганди	**Regression**	**SVM**	**Random Forest**	**Naive Bayes**
Апеляція до авторитету	0.59	**0.67**	0.54	0.60
Помилка хибної дихотомії	0.61	0.62	0.55	**0.64**
Зведення до Гітлера	**0.69**	0.63	0.62	0.58
Червоний оселедець	**0.69**	0.64	0.58	0.61
Гасла	0.62	**0.63**	0.56	0.62
Кліше, що припиняють думання	0.62	**0.68**	0.62	0.61
Ватабоутізм	0.64	**0.66**	0.58	0.56

Для решти прийомів пропаганди SMOTE балансування не застосовувалось; результати застосування традиційного підходу машинного навчання наведено у таблиці 4.8.

Результати з таблиці 4.8 також коливаються від 0.6 до 0.67, що має схожий результат із застосуванням SMOTE для балансування (таблиця 4.7). Однак отримані результати також не є задовільними. Ілюстрація експерименту щодо застосування традиційного підходу машинного навчання наведена на рисунку 4.9.

Таблиця 4.8

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди за метрикою Accuracy

Прийоми пропаганди	**Regression**	**SVM**	**Random Forest**	**Naive Bayes**
Апеляція до страху	0.62	0.61	0.55	0.59
Причинно-надмірна спрощеність	0.58	0.62	0.59	0.55
Сумнів	0.6	0.58	0.63	0.59
Перебільшення	0.61	0.63	0.61	0.53
Розмахування прапором	0.62	0.61	0.64	0.6
Навішування ярликів	0.67	0.62	0.61	0.61
Заряджена мова	0.6	0.59	0.6	0.58
Мінімізація	0.62	0.61	0.61	0.60
Name Calling	0.55	0.59	0.57	0.61
Повторення	0.69	0.7	0.61	0.55

Відповідно до таблиці 4.8, жоден із підходів не досягає високої точності, що вказує на обмеження традиційних моделей у задачах виявлення прийомів пропаганди. Це свідчить про необхідність розгляду інших методів ШІ, здатних краще інтерпретувати тонкі мовні конструкції, притаманні прийомам пропаганди.

Наступним експериментом проведено дослідження застосування рекурентних нейронних мереж із порівнянням використань 3-х видів архітектур: RNN, LSTM та GRU. Дані експерименту без використання SMOTE балансування наведено у таблиці 4.9.

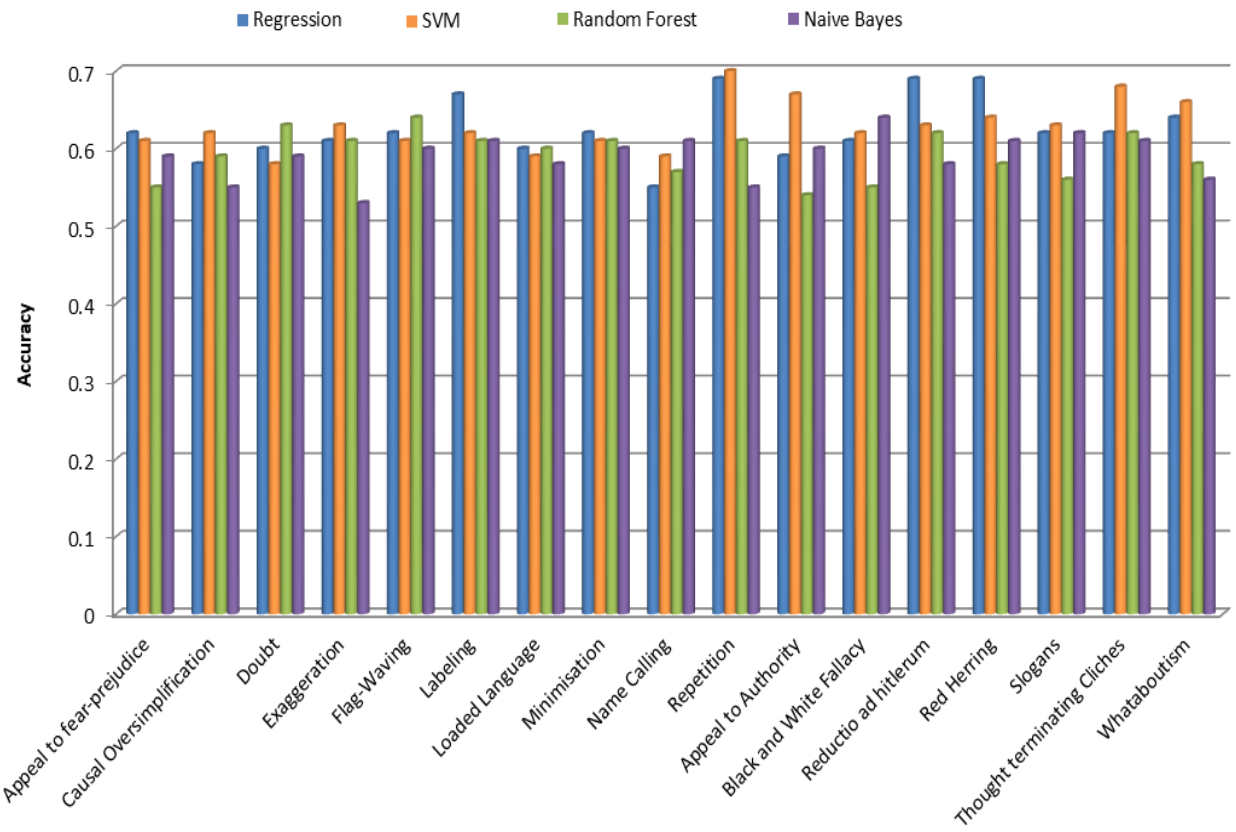


Рис. 4.9 – Порівняння точності моделей для традиційного підходу машинного навчання

Відповідно до даних таблиці 4.9 результати для усіх прийомів пропаганди, окрім «Кліше, що припиняють думання», є вищими та знаходяться у діапазоні від 0.66 до 0.8. Однак до даного прийому пропаганди в подальшому застосовано SMOTE балансування, що може дозволити покращити показник.

Наступним експериментом буде використання SMOTE балансування до навчання нейромережевих моделей для «Апеляція до авторитету», «Помилка хибної дихотомії», «Зведення до Гітлера», «Червоний оселедець», «Гасла», «Кліше, що припиняють думання» та «Ватабоутізм».

Таблиця 4.9

Підхід на основі рекурентних нейромереж для виявлення прийомів пропаганди за метрикою Accuracy

Прийоми пропаганди	**RNN**	**LSTM**	**GRU**
Апеляція до страху	0.69	0.69	**0.71**
Причинно-надмірна спрощеність	0.73	0.71	**0.76**
Сумнів	**0.75**	0.7	0.74
Перебільшення	0.64	0.72	**0.75**
Розмахування прапором	0.69	0.7	**0.79**
Навішування ярликів	0.69	0.73	**0.8**
Заряджена мова	**0.71**	0.7	0.68
Мінімізація	**0.78**	**0.78**	0.74
Обзивання	**0.76**	0.74	**0.76**
Повторення	0.74	0.75	**0.76**
Апеляція до авторитету	0.71	0.72	**0.73**
Помилка хибної дихотомії	0.7	0.68	**0.72**
Зведення до Гітлера	**0.75**	0.68	0.71
Червоний оселедець	0.65	**0.72**	0.70
Гасла	0.74	0.68	**0.75**
Кліше, що припиняють думання	0.63	**0.66**	0.65
Ватабоутізм	0.67	**0.69**	**0.69**

Результати наведені в таблиці 4.10 свідчать, що використання SMOTE балансування дає позитивний ефект на точність виявлення прийомів пропаганди.

Таблиця 4.10

Підхід на основі рекурентних нейромереж для виявлення прийомів пропаганди за метрикою Accuracy із SMOTE балансуванням

Прийоми пропаганди	**RNN**	**LSTM**	**GRU**
Апеляція до авторитету	0.7	0.72	**0.73**
Помилка хибної дихотомії	**0.72**	0.7	**0.72**
Зведення до Гітлера	0.73	**0.74**	**0.74**
Червоний оселедець	0.69	0.73	**0.75**
Гасла	0.72	**0.76**	0.72
Кліше, що припиняють думання	0.69	0.76	**0.78**
Ватабоутізм	0.68	0.7	**0.78**

Не вдалося покращити виявлення прийому «Зведення до Гітлера», де результати до SMOTE балансування були на 0.01 вище, а також прийоми «Апеляція до авторитету» та «Помилка хибної дихотомії» залишились на тому ж рівні, що і до виконання балансування.

На рисунку 4.10 показано порівняння найвищих отриманих показників точності для рекурентних нейромережевих моделей.

Останнім етапом дослідження є застосування підходу на основі використання моделей-трансформерів, що містить порівняння BERT-подібних моделей: RoBERTa, BERT, ELECTRA. Були використані попередньо натреновані моделі з ресурсу Hugging Face [153], які були донавчені вищеописаним способом протягом 3-х епох навчання. Отримані результати без використання SMOTE балансування наведені у таблиці 4.11.

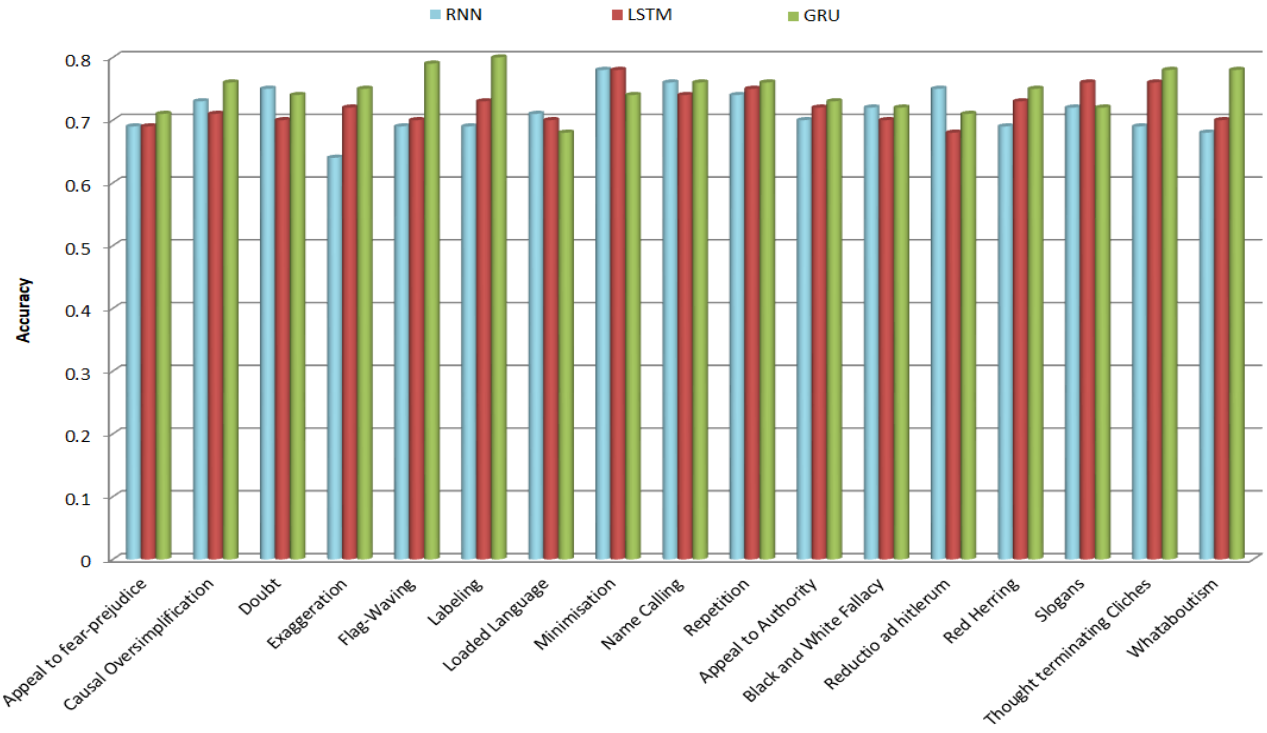


Рис. 4.10 – Порівняння точності моделей для підходу на основі рекурентнх нейронних мереж

Із даних таблиці 4.11 видно, що BERT-подібні нейромережеві архітектури значно краще виявляють прийоми пропаганди порівняно з рекурентними та традиційним підходами до навчання. Це пояснюється тим, що такі архітектури є контекстно-орієнтованими, що є важливим аспектом для виявлення прийомів пропаганди.

Застосування SMOTE для балансування дозволило підвищити точність виявлення прийому «Червоний оселедець» нейромережевою моделлю Ukr-Electra-Base до 0.89, а також прийому «Ватабоутізм» до 0.83 з використанням архітектури Bert-Base-Multilingual-Cased.

Порівняння найвищих оцінок за метрикою точності для 3-х розглянутих підходів наведено на рис. 4.11.

Таблиця 4.11

Підхід на основі трансформерних моделей для виявлення прийомів пропаганди

Прийоми пропаганди	**bert-base-multilingual-cased**	**roberta-base**	**ukr-electra-base**
Апеляція до страху	0.81	0.8	**0.87**
Причинно-надмірна спрощеність	0.78	0.79	**0.82**
Сумнів	**0.93**	0.9	0.87
Перебільшення	**0.8**	**0.8**	**0.8**
Розмахування прапором	**0.92**	0.9	0.89
Навішування ярликів	**0.96**	0.94	**0.96**
Заряджена мова	0.93	**0.97**	0.94
Мінімізація	0.89	0.86	**0.9**
Обзивання	**0.92**	**0.92**	0.91
Повторення	0.93	**0.94**	**0.94**
Апеляція до авторитету	0.87	**0.89**	0.88
Помилка хибної дихотомії	0.89	**0.91**	0.88
Зведення до Гітлера	0.85	**0.87**	0.86
Червоний оселедець	0.67	**0.8**	0.78
Гасла	0.84	**0.86**	0.83
Кліше, що припиняють думання	**0.8**	0.73	0.79
Ватабоутізм	**0.79**	0.78	0.78

Відповідно до рис. 4.14 традиційний підхід на основі машинного навчання очікувано показав гірші результати, оскільки він не вміє бачити контекст, що є важливим для виявлення прийомів пропаганди.

Рекурентні нейромережеві моделі хоч і показали результати вищі від традиційного підходу, однак все ще мають проблеми з обробкою довгих

залежностей. Найвищі результати з проведеного експерименту виявились у підході на основі моделей-трансформерів, що пояснюється використовуваними механізмами самоуваги, які дозволяють кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами та ефективно охоплювати довготривалі залежності, що характерні для проявів прийомів пропаганди.

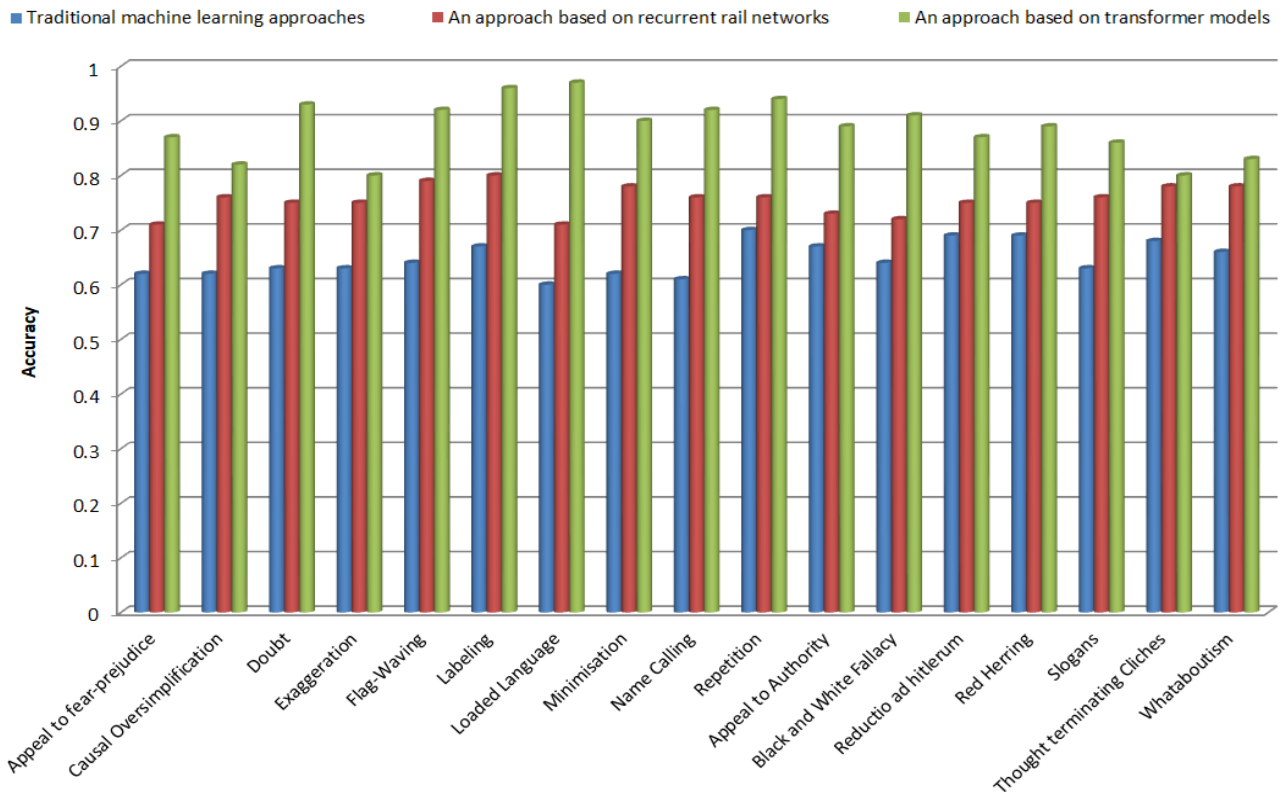


Рис. 4.11 – Порівняння точності моделей альтернативних підходів для виявлення прийомів пропаганди

Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 79,03% (мінімальні значення точності отримані для методики «Ватабоутізм»), що краще за відомі аналоги [41] щодо виявлення пропаганди незалежно від використовуваних методик.

Порівняно з відомими аналогами [28] підвищилась точність виявлення різних пропагандистських прийомів:

– для прийому «Апеляція до авторитету» точність виявлення зросла на 9.81% (існуючий метод 77.27%, розроблений метод 87.08%);

– для прийому «Причинно-надмірна спрощеність» точність виявлення зросла на 11.99% (існуючий метод 70.1%, розроблений метод 82.09%);

– для прийому «Сумнів» точність виявлення зросла на 75.32% (існуючий метод 17.78%, розроблений метод 93.1%);

– для прийому «Перебільшення» точність виявлення зросла на 26.04% (існуючий метод 54.17%, розроблений метод 80.21%);

– для прийому «Розмахування прапором» точність виявлення зросла на 27.65% (існуючий метод 64.52%, розроблений метод 92.17%);

– для прийому «Навішування ярликів» точність виявлення зросла на 48.57% (існуючий метод 47.43%, розроблений метод 96.0%);

– для прийому «Заряджена мова» точність виявлення зросла на 42.9% (існуючий метод 54.17%, розроблений метод 97.07%);

– для прийому «Обзивання» точність виявлення зросла на 44.6% (існуючий метод 47.43%, розроблений метод 92.03%);

– для прийому «Повторення» точність виявлення зросла на 58.11% (існуючий метод 35.98%, розроблений метод 94.09%);

– для прийому «Апеляція до авторитету» точність виявлення зросла на 11.84% (існуючий метод 77.18%, розроблений метод 89.02%);

– для прийому «Помилка хибної дихотомії» точність виявлення зросла на 36.68% (існуючий метод 54.55%, developed method 91.23%);

– для прийому «Зведення до Гітлера» точність виявлення зросла на 62.31% (існуючий метод 25.0%, розроблений метод 87.31%);

– для прийому «Червоний оселедець» точність виявлення зросла на 41.07% (існуючий метод 39.22%, розроблений метод 80.29%);

– для прийому «Гасла» точність виявлення зросла на 10.54% (існуючий метод 75.5%, розроблений метод 86.04%);

– для прийому «Кліше, що припиняють думання» точність виявлення зросла на 26.74% (існуючий метод 53.57%, розроблений метод 80.31%);

– для прийому «Ватабоутізм» точність виявлення зросла на 39.81% (існуючий метод 39.22%, розроблений метод 79.03%).

Наступна частина дослідження присвячена визначенню впливу множини семантичних маркерів на здатність нейромереж архітектури трансформер до класифікації прийомів пропаганди.

На рис. 4.12 наведено приклад ідентифікації прийомів пропаганди та відповідні оцінки за множиною маркерів. Запропонований метод дозволяє отримати додаткову поясненість, порівнюючи отримані результати із таблицею 3.2. До прикладу, при аналізі контенту з рис. 4.12, оригінальний україномовний текст якого: «*Українська держава опинилася на межі катастрофи! Всі ми можемо стати свідками кінця нашої незалежності. Ворог на кожному кроці, і якщо ми не діятимемо негайно, нас усіх чекає жахлива доля. Вони не просто хочуть забрати нашу землю - вони прагнуть знищити нашу культуру, мову, ідентичність. Наші вороги вже перебувають всередині країни, приховані серед нас, і готові завдати удару в будь-який момент. Ми маємо об'єднатися і боротися, інакше все буде втрачено! Вчорашні події на кордоні - це лише початок. Сотні тисяч військових збираються на наших кордонах, готуючи наступ. Ворог намагається залякати нас своїми погрозами, і ми не можемо дозволити собі бути слабкими. Нам потрібні всі ресурси, вся наша міць і відвага, щоб протистояти цій загрозі. Вони перебільшують свої сили, але це не повинно нас заспокоювати - навпаки, це має підштовхнути нас до дій*.», було

визначено застосування прийомів пропаганди «Апеляція до страху» з оцінкою 0.79 та «Перебільшення» з оцінкою 0.68.

Рис. 4.12 – Приклад аналізу виявлених прийомів пропаганди розробленим програмним забезпеченням

Дані прийоми пропаганди (відповідно до таблиці 3.2) характеризуються усередненими значеннями семантичних маркерів: «Емоційність тексту» – 0.7, «Булінг» – 0.5, «Страх» – 0.8, «Мова ворожнечі» – 0.7 для прийому «Апеляція до страху». Водночас усереднено-характерними для прийому «Перебільшення» значеннями маркерів є «Емоційність тексту» – 0.8, «Булінг» – 0.4, «Страх» – 0.7, «Мова ворожнечі» – 0.6.

Відповідно до значень, отриманих програмним продуктом, результати корелюють із значеннями таблиці 3.2 із незначними розбіжностями. Розбіжності

також пояснюються із оцінкою співвіднесення тестового тексту до еталона, адже прийом «Апеляція до страху» оцінений на 0.79, отже і нейромережа має деякі сумніви (у даному прикладі трохи завищено семантичну ознаку «Емоційність тексту» 0.9 порівняно з усередненим значенням 0.7) та «Перебільшення» з оцінкою 0.68 (усі семантичні ознаки від еталона мають розбіжність від 0.05 до 0.15).

Приклад візуальної поясненості щодо виявлення прийому пропаганди «Повторення» наведено на рис. 4.13 (використано оригінальний текст повсякденною українською мовою із збереженням орфографії та помилок).

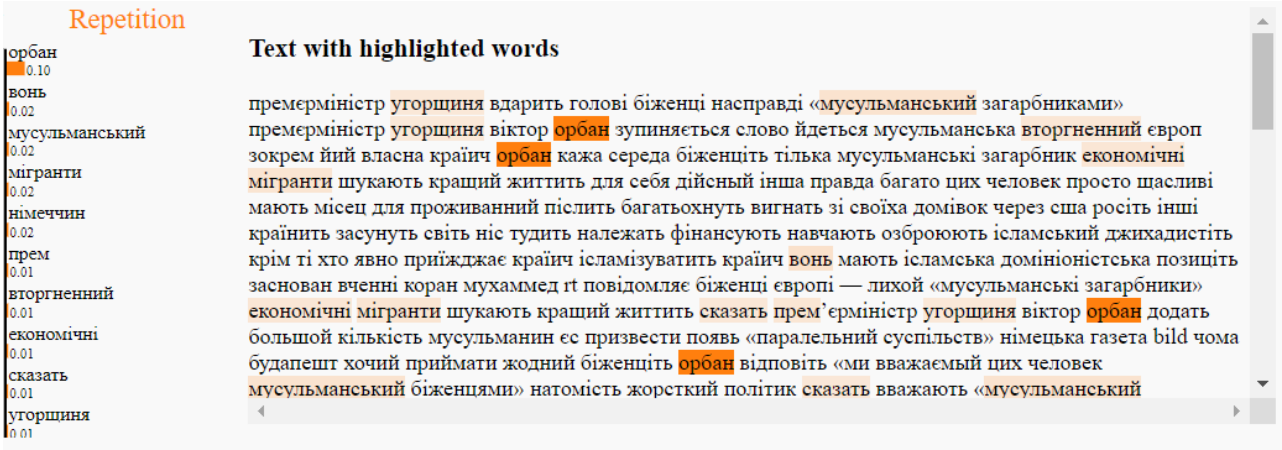


Рис. 4.13 – Візуальна аналітика щодо виявлення прийому пропаганди «Повторення» розробленим програмним забезпеченням

Відповідно до рис. 4.13 є багаторазові повторення фраз на кшталт «економічні мігранти», «мусульманський», «Орбан» тощо, що корелює з означенням прийому пропаганди «Повторення», яке є повторенням якогось повідомлення знову і знову, щоб суб'єкти, на яких спрямовано вплив, зрештою прийняли його. Отже, запропонований метод дозволяє ефективно виявляти прийоми пропаганди та має перевагу у точності порівняно із запропонованими моделями, які використовують підхід багатокласової класифікації.

Порівняння застосування множини семантичних маркерів з існуючими аналогами та попередніми найкращими результатами власних досліджень, наведених у таблиці 4.11, подано у таблиці 4.12.

Таблиця 4.12

Порівняння запропонованих та існуючих підходів до класифікації пропаганди

Прийоми пропаганди	**Accuracy**		
	Існуючі аналоги [27]	**Трансформери без застосування множини семантичних маркерів**	**Трансформер із застосуванням множини семантичних маркерів**
Апеляція до страху	0.77	0.87	**0.88** (+0.01)
Причинно-надмірна спрощеність	0.7	0.82	**0.83** (+0.01)
Сумнів	0.17	0.93	0.91 (-0.02)
Перебільшення	0.54	0.8	**0.82** (+0.02)
Розмахування прапором	0.64	0.92	0.91 (-0.01)
Навішування ярликів	0.47	0.96	0.93 (-0.03)
Заряджена мова	0.54	0.97	0.97 (0.00)
Мінімізація	-	0.9	**0.91** (+0.01)
Обзивання	0.47	0.92	0.9 (-0.02)
Повторення	0.36	0.94	**0.95** (+0.01)
Апеляція до авторитету	0.77	0.89	**0.9** (+0.01)
Помилка хибної дихотомії	0.54	0.91	0.91 (0.00)
Зведення до Гітлера	0.25	0.87	0.87 (0.00)
Червоний оселедець	0.39	0.8	**0.89** (+0.09)
Гасла	0.76	0.86	0.85 (-0.01)
Кліше, що припиняють думання	0.53	0.8	**0.83** (+0.03)
Ватабоутізм	0.39	0.79	**0.83** (+0.04)

Результати порівняння розроблених та існуючих підходів до класифікації прийомів пропаганди з таблиці 4.12 відображені на рис. 4.14.

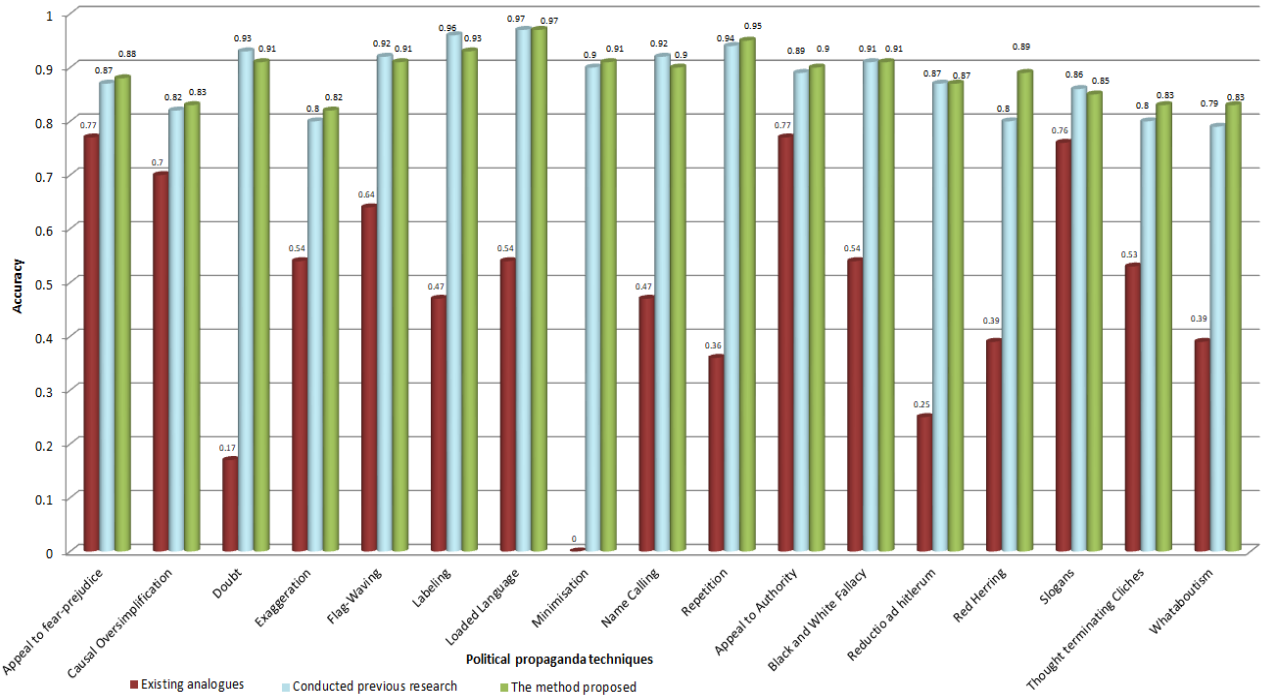


Рис. 4.14 – Порівняння розроблених та існуючих підходів до класифікації прийомів пропаганди

Відповідно до даних таблиці 4.12 та рис. 4.14 для більшості прийомів пропаганди застосування доповненої множини семантичних маркерів дає приріст у класифікації. Також варто зауважити, що у колонці трансформери без застосування множини семантичних маркерів таблиці 4.12 взяті найкращі сукупні результати по всіх моделях (bert-base-multilingual-cased, roberta-base та ukr-electra-base), водночас значення у колонці «трансформер» із застосуванням множини семантичних маркерів взято тільки за даними моделі bert-base-multilingual-cased [134].

Незначне покращення точності (на 0.01) спостерігається для кількох пропагандистських прийомів: «Апеляція до страху», «Причинно-надмірна

спрощеність», «Мінімізація», «Повторення», «Апеляція до авторитету». Прийом «Апеляція до страху» показав незначне підвищення точності з 0.87 до 0.88, а класифікація прийому «Причинно-надмірна спрощеність» зросла за метрикою Accuracy з 0.82 до 0.83, Accuracy для прийому «Мінімізація» показала зростання з 0.90 до 0.91, а «Повторення» – з 0.94 до 0.95. Також прийом «Апеляція до авторитету» показав підвищення точності з 0.89 до 0.90. Ці результати свідчать про ефективність розробленого методу в підвищенні точності виявлення прийомів пропаганди.

Декілька прийоми пропаганди показали помітні покращення. Прийом «Перебільшення» показав підвищення точності з 0.80 до 0.82, прийом «Червоний оселедець» показав значне покращення з 0.80 до 0.89 (зростання точності на 0.09), прийом «Ватабоутізм» – з 0.79 до 0.83, а прийом пропаганди «Кліше, що припиняють думання» має приріст за метрикою точності з 0.80 до 0.83.

Та для деяких прийомів спостерігається зниження точності. Наприклад, для прийому пропаганди «Навішування ярликів» точність знизилась з 0.96 до 0.93, для прийому «Обзивання» – з 0.92 до 0.90. Точність прийому «Гасла» знизилась з 0.86 до 0.85, а «Сумнів» – з 0.93 до 0.91. Це свідчить про необхідність подальшої оптимізації методів виявлення цих конкретних прийомів пропаганди.

Також є ряд прийомів, які зберегли свою точність без змін: «Заряджена мова» залишилася на рівні 0.97, «Помилка хибної дихотомії» – на рівні 0.91, а «Зведення до Гітлера» – на рівні 0.87. Це свідчить про стабільність запропонованого методу у виявленні цих прийомів.

Отримані результати свідчать про спроможність запропонованого методу ефективно виявляти прийоми пропаганди, а також візуально оцінювати, які саме дані вплинули на рішення моделей щодо наявних прийомів пропаганди у користувацькому тексті.

Приклад поясненості з використанням моделі LIME щодо наявності у тексті прийому «Апеляція до страху» наведено на рис. 4.15.

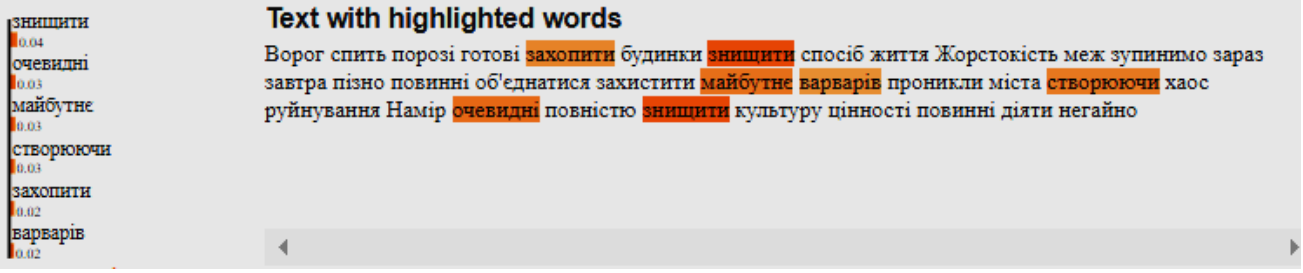


Рис. 4.15 – Приклад пояснення результатів щодо наявності у тексті прийому «Апеляція до страху»

Суть прийому «Апеляція до страху» полягає у створенні або підсиленні відчуття загрози та страху з метою змусити людей прийняти певні ідеї або дії, які вважаються захисними або необхідними для уникнення небезпеки. Відповідно до рис. 4.15 вагомими словами, що вплинули на рішення моделі щодо присутності прийому «Апеляція до страху» є такі слова, як «захопити», «знищити», «майбутнє», «варварів» тощо, що цілком відповідає визначенню даного прийому.

Однак усереднено точність зросла на 0.361 при реалізації методу без множини семантичних маркерів та на 0.368 при реалізації методу з множиною семантичних маркерів порівняно з [28]. Реалізація з використанням множини семантичних маркерів дозволяє усереднено збільшити точність на 0.007 порівняно з власною реалізацією, що не враховує додавання такої множини. Застосування методу з додатковою множиною маркерів дає можливість не лише підняти загальну точність ідентифікації, а й отримати додаткову поясненість рішень за допомогою порівняння отриманих значень з таблицею еталонних.

Отже, запропонований метод виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень, що ґрунтується на

використанні набору нейромереж окремих для кожного прийому пропаганди, що навчаються на модифікованих розмічених даних з доповненою множиною маркерів, дозволяє виявляти прийоми пропаганди з середньою точністю 0.886, що, зважаючи на здатність пропаганди маскуватись у контекстах повідомлень та новин, є високим показником.

Варто зазначити, що з огляду на вищенаведені дослідження існує потенціал подальших вишукувань, які можуть бути спрямовані на розширення датасету, а також на розширення множини маркерів та навчання відповідних моделей машинного навчання.

4.4. Дослідження методу виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень

Для дослідження методу виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання було розроблене експериментальне програмне забезпечення у вигляді інтелектуальної системи, архітектура якої наведена на рис. 4.16. Вона складається із модуля попередньої обробки тексту, модуля виявлення прийомів пропаганди, модуля пошуку NER, модуля розширення множини об’єктів пропаганди семантично-близькими представленнями, модуля побудови контекстних вікон та модуля визначення об’єктів пропаганди.

Модуль пошуку NER призначений для знаходження всіх іменованих сутностей документа до його попередньої обробки. Модуль використовує нейромережеву бібліотеку «Stanza».

Модуль попередньої обробки тексту включає видалення стоп-слів, стоп-символів та лематизацію. Його функціональність використовується модулем

розширення множини об’єктів пропаганди семантично-близькими представленнями та модулем виявлення прийомів пропаганди.

Рис. 4.16 – Програмна архітектура інтелектуальної системи

Модуль розширення множини об’єктів пропаганди семантично-близькими представленнями призначений для пошуку семантично-близьких об’єктів до знайдених NER. Для його реалізації використовується бібліотека «FastText». Семантично близькі представлення підбираються для лематизованого тексту, щоб уникати повторів.

Модуль визначення об’єктів пропаганди агрегує в собі результати роботи модулів пошуку NER та розширення множини об’єктів пропаганди семантично-близькими представленнями.

Модуль виявлення прийомів пропаганди призначений для пошуку застосованих у тексті прийомів пропаганди. Працює за допомогою використання навчених нейромережевих моделей, які натреновані окремо на ідентифікацію кожного з досліджуваних прийомів пропаганди («Апеляція до страху», «Причинно-надмірна спрощеність», «Сумнів», «Перебільшення», «Розмахування прапором», «Навішування ярликів», «Заряджена мова», «Мінімізація», «Обзивання», «Повторення», «Апеляція до авторитету», «Помилка хибної дихотомії», «Зведення до Гітлера», «Червоний оселедець», «Гасла», «Кліше, що припиняють думання», «Ватабоутізм»).

Модуль побудови контекстних вікон виділяє речення, в яких зустрічаються знайдені об’єкти пропаганди. Речення в межах одного об’єкта не дублюються. Отримані контекстні вікна для NER та їх розширеної множини далі проходять повторну перевірку через модуль виявлення прийомів пропаганди, який і дає відповідь, до якого прийому належать об’єкти в рамках контекстного вікна.

Вихідними даними є визначені прийоми пропаганди; визначені об’єкти пропаганди; визначені приналежності прийомів до об’єктів.

Дослідження методу виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень [154] полягає у послідовному порівнянні даних, отриманих розробленим програмним забезпеченням [155], та оцінок експертів. На рисунку 4.17 наведено покрокову ілюстрацію запропонованого підходу.

До прикладу, в межах тексту *«Якщо ви не приєднаєтеся до нашої партії, ви будете приречені на страждання, як жителі Лівії. Відмова від нашого плану може призвести до катастрофічних наслідків, подібних до тих, що сталися в*

Україні під час конфлікту. Влада, яка нині намагається контролювати країну, уже показала свою неспроможність. Чи готові ви ризикувати безпекою своєї родини через ці некомпетентні рішення?») першим кроком є пошук NER.

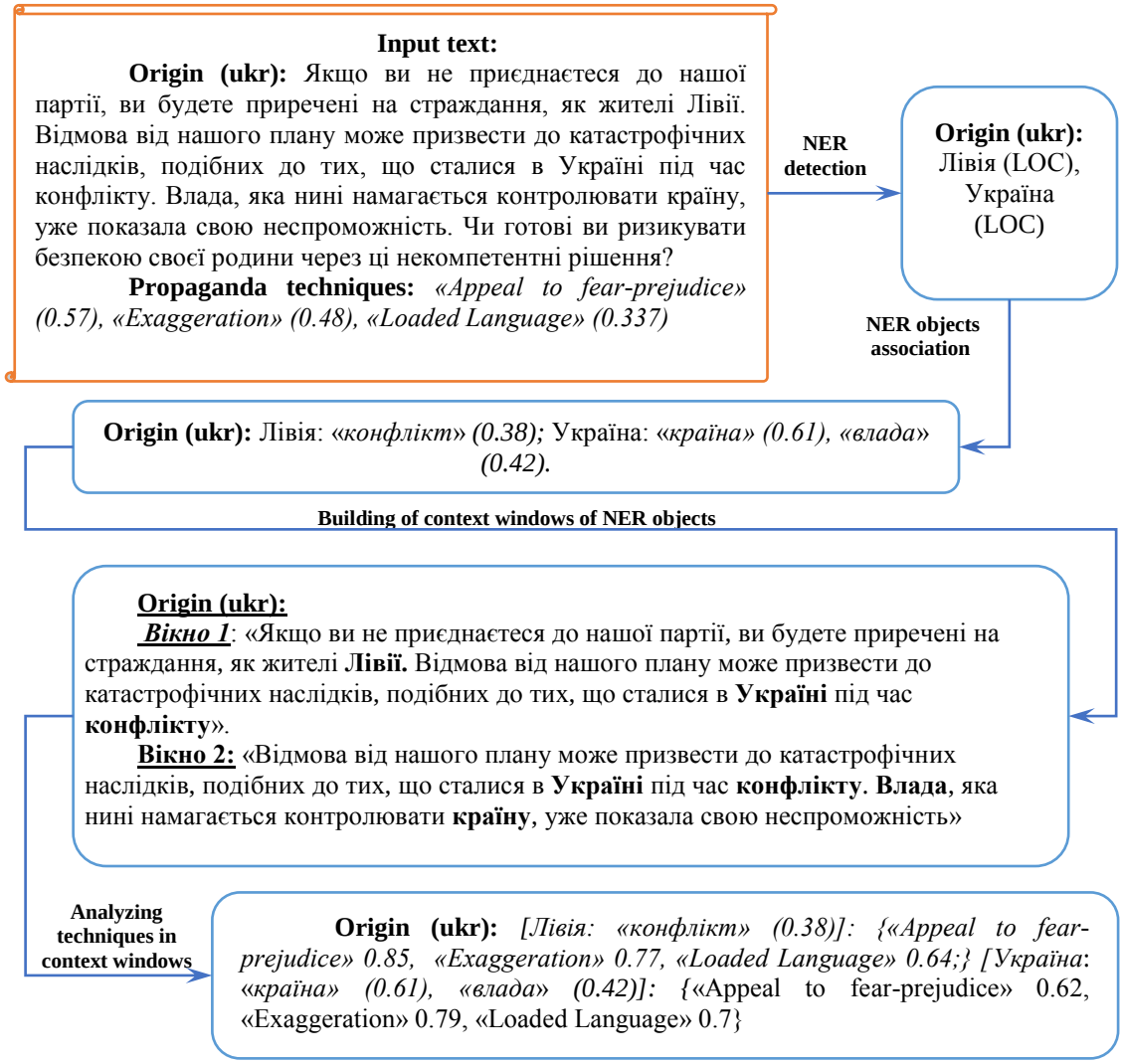


Рис. 4.17 – Приклад перетворення даних за пошуком об’єктів пропаганди розробленим методом

У даному тексті знайдено 2 іменовані сутності: «*Лівія*» (*LOC*) та «*Україна*» (*LOC*). Множину знайдених NER далі необхідно розширити шляхом підбору семантично близьких об’єктів.

Після проходження етапу NER множина доповниться і для даного прикладу стане така: {«*Лівія*»: «*конфлікт*» (0.38)}; {«*Україна*»: «*країна*» (0.61), «*влада*» (0.42)}.

Наступним етапом є побудова контекстних вікон для знайдених об'єктів. Вікно 1. {Лівія: «конфлікт» (0.38)}: *«Якщо ви не приєднаєтеся до нашої партії, ви будете приречені на страждання, як жителі Лівії. Відмова від нашого плану може призвести до катастрофічних наслідків, подібних до тих, що сталися в Україні під час конфлікту».* Вікно 2. *{Україна: «країна» (0.61), «влада» (0.42)}*: *«Відмова від нашого плану може призвести до катастрофічних наслідків, подібних до тих, що сталися в Україні під час конфлікту. Влада, яка нині намагається контролювати країну, уже показала свою неспроможність».*

Останнім кроком є пошук прийомів у контекстних вікнах. Оцінка наявності прийому здійснюється шляхом оцінки контекстного вікна нейромережею, що навчена ідентифікувати виражені у тексті прийоми. Для даного прикладу: [*Лівія: «конфлікт» (0.38)*] прийом «Апеляція до страху» оцінено на 0.85, «Перебільшення» – на 0.77 і «Заряджена мова» – на 0.64. Водночас для: [*Україна: «країна» (0.61), «влада» (0.42)*] прийом «Апеляція до страху» оцінено на 0.62, «Перебільшення» – на 0.79 і «Заряджена мова» – на 0.7.

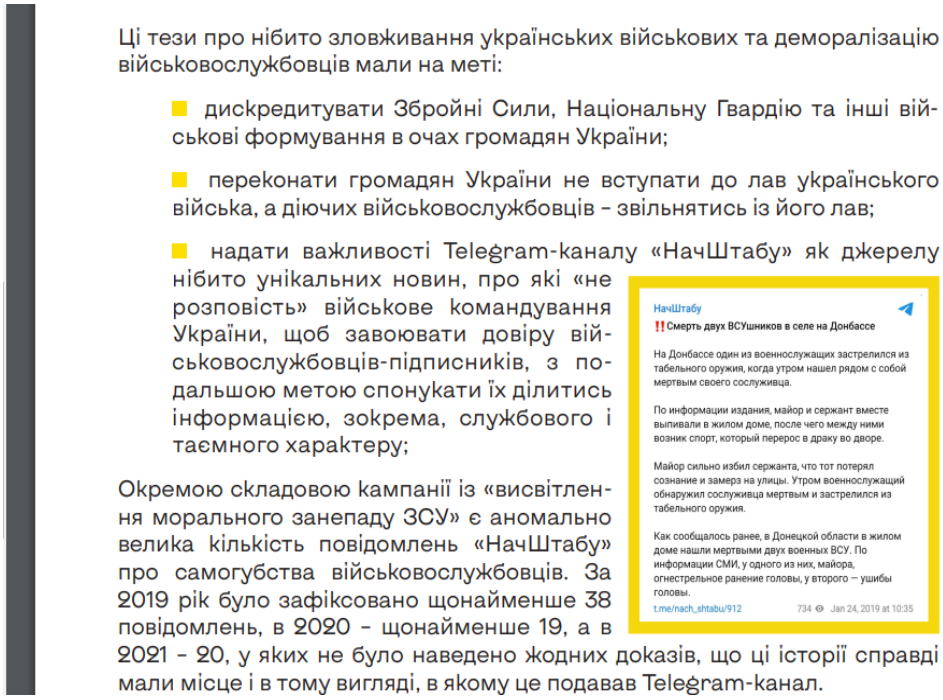
Враховуючи вміст тестового тексту, тут майже рівномірно розподілені прийоми по контекстних вікнах. Однак для першого контекстного вікна менш вираженим є прийом «Заряджена мова» та більш вираженим прийом «Апеляція до страху» порівняно із другим контекстним вікном.

Дослідження ефективності запропонованого підходу для виявлення прийомів пропаганди та об'єктів, на які вони спрямовані, свідчить що результати корелюють з висновками дослідників із «Центру стратегічних комунікацій» [146]. До прикладу, пост із пропагандистського каналу «НачШтабу» (рис. 4.18) та отриманий експертом висновок.

Відповідно до висновку експерта (рис. 4.18) тези про нібито зловживання українських військових та деморалізацію військовослужбовців мали на меті:

– дискредитувати Збройні Сили, Національну Гвардію та інші військові формування в очах громадян України;

– переконати громадян України не вступати до лав українського війська, а діючих військовослужбовців – звільнятись із його лав.


Ці тези про нібито зловживання українських військових та деморалізацію військовослужбовців мали на меті:

- дискредитувати Збройні Сили, Національну Гвардію та інші військові формування в очах громадян України;
- переконати громадян України не вступати до лав українського війська, а діючих військовослужбовців – звільнятись із його лав;
- надати важливості Telegram-каналу «НачШтабу» як джерелу нібито унікальних новин, про які «не розповість» військове командування України, щоб завоювати довіру військовослужбовців-підписників, з подальшою метою спонукати їх ділитись інформацією, зокрема, службового і таємного характеру;

НачШтабу

‼️Смерть двух ВСУшников в селе на Донбассе

На Донбассе один из военнослужащих застрелился из табельного оружия, когда утром нашел рядом с собой мертвым своего сослуживца.

По информации издания, майор и сержант вместе выпивали в жилом доме, после чего между ними возник спорт, который перерос в драку во дворе.

Майор сильно избил сержанта, что тот потерял сознание и замерз на улицы. Утром военнослужащий обнаружил сослуживца мертвым и застрелился из табельного оружия.

Как сообщалось ранее, в Донецкой области в жилом доме нашли мертвыми двух военных ВСУ. По информации СМИ, у одного из них, майора, огнестрельное ранение головы, у второго – ушибы головы.

t.me/nach_shtabu/912 734 Jan 24, 2019 at 10:35

Окремою складовою кампанії із «висвітлення морального занепаду ЗСУ» є аномально велика кількість повідомлень «НачШтабу» про самогубства військовослужбовців. За 2019 рік було зафіксовано щонайменше 38 повідомлень, в 2020 - щонайменше 19, а в 2021 - 20, у яких не було наведено жодних доказів, що ці історії справді мали місце і в тому вигляді, в якому це подавав Telegram-канал.

Рис. 4.18 – Аналіз тексту, що містить пропаганду «Центром стратегічних комунікацій» [146]

У розрізі аналізу отриманого за допомогою використання розробленої програмної установки отримано такі дані:

– використані прийоми пропаганди з силами прояву: «Заряджена мова» з оцінкою 0.582 та «Повторення» з оцінкою 0.317;

– об’єкти пропаганди: NER разом з семантично близькими до них словами з оцінками близькості (ЗСУ [вранці, вулиця, суперечка, виникнути]; донбас [зма, випивати, разом, раніше]; донецький область [вони, виникнути]);

– оцінки відповідності об’єктів пропаганди до використаних прийомів (ЗСУ [Заряджена мова 0.593; Повторення 0.612], донбас [Заряджена мова 0.407; Повторення 0.35], донецький область [Заряджена мова 0.361; Повторення 0.71]);

– візуальне подання знайдених об’єктів у користувацькому тексті.

Використання прийому «Заряджена мова» використовується в тексті для опису конфліктів і насильства, наприклад, «сильно побив», «замерз на вулиці», «застрелився з табельної зброї». Це відповідає меті дискредитації Збройних Сил та висвітлення їх у негативному світлі, що відповідає висновку експерта.

Використання прийому «Повторення» використовується у вигляді повторення інформації про смерть військовослужбовців і насильницьких дій. Повторення допомагає підсилити негативний вплив і зміцнити негативне враження. Це відповідає меті переконати громадян не вступати до лав війська, а діючих військовослужбовців – звільнятись. Приклад виконаного аналізу програмою наведено на рис. 4.19.

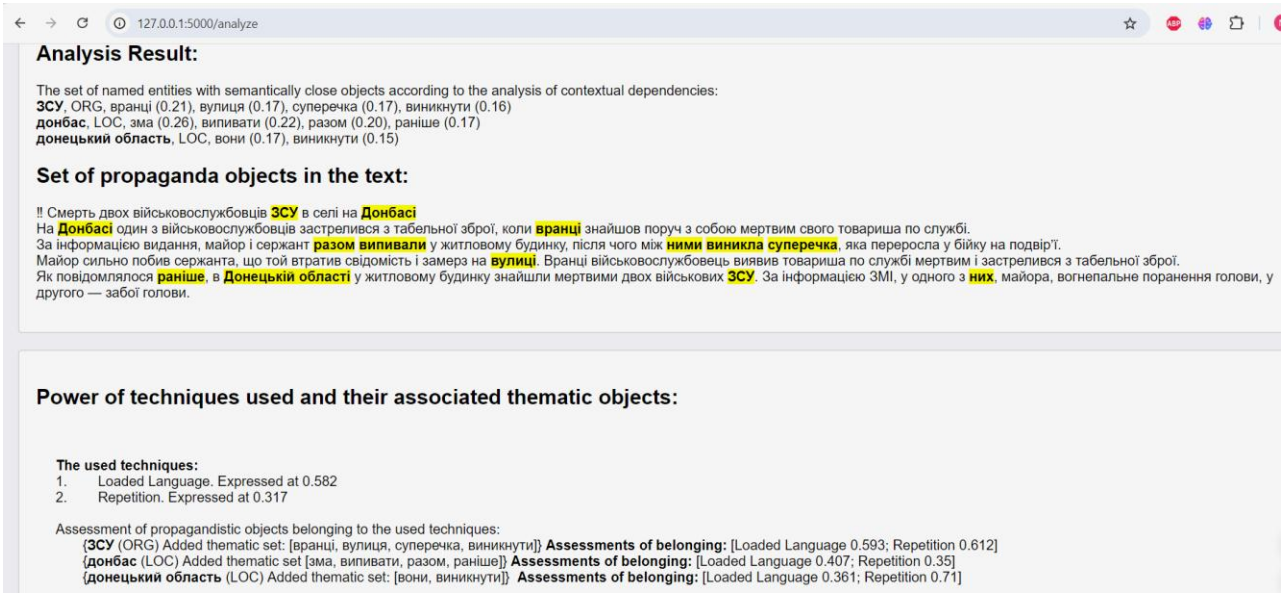


Рис. 4.19 – Приклад аналізу тексту розробленим програмним забезпеченням

Відповідно до рис. 4.19 було детектовано не лише об’єкти пропаганди та використані прийоми, а й визначено приналежність знайдених об’єктів до використаних прийомів.

Також результати дослідження методу виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень співвідносяться з висновками дослідників «Центру протидії дезінформації» [156]. З їх звіту було взято текстовий допис та досліджено створеним програмним забезпеченням. Наведені висновки експертів та досліджуваний текст на рис. 4.20.

Рис. 4.20 – Аналіз тексту з пропагандою дослідниками «Центру протидії дезінформації» [156]

Відповідно до висновків експертів пост спрямовано на:

– виправдання використання ядерної зброї проти США;
– подальше втручання Заходу загрожує загостренню війни;

– виставлення Заходу як слабкого суперника (дискредитація).

У межах аналізу, отриманого за допомогою розробленого програмного забезпечення, були отримані наступні дані (рис. 4.21):

– використано прийоми пропаганди з силами прояву: «Апеляція до страху» з оцінкою 0.598, «Заряджена мова» з оцінкою 0.46 та «Перебільшення» з оцінкою 0.356;

– об'єкти пропаганди: NER разом із семантично близькими до них словами з оцінками близькості, вихідні дані: «США [зброя, американський, перше, розуміти]; ППО [територія, той, північ, церемонитися, острів]»;

– оцінка важливих зв'язків об'єктів пропаганди з прийомами, що використовується, програмними вихідними даними: «США [Апеляція до страху 0.578; Заряджена мова 0.473; Перебільшення 0.311], ППО [Апеляція до страху 0.611; Заряджена мова 0.4; Перебільшення 0.379]»;

– візуальне представлення знайдених об'єктів у користувацькому тексті.

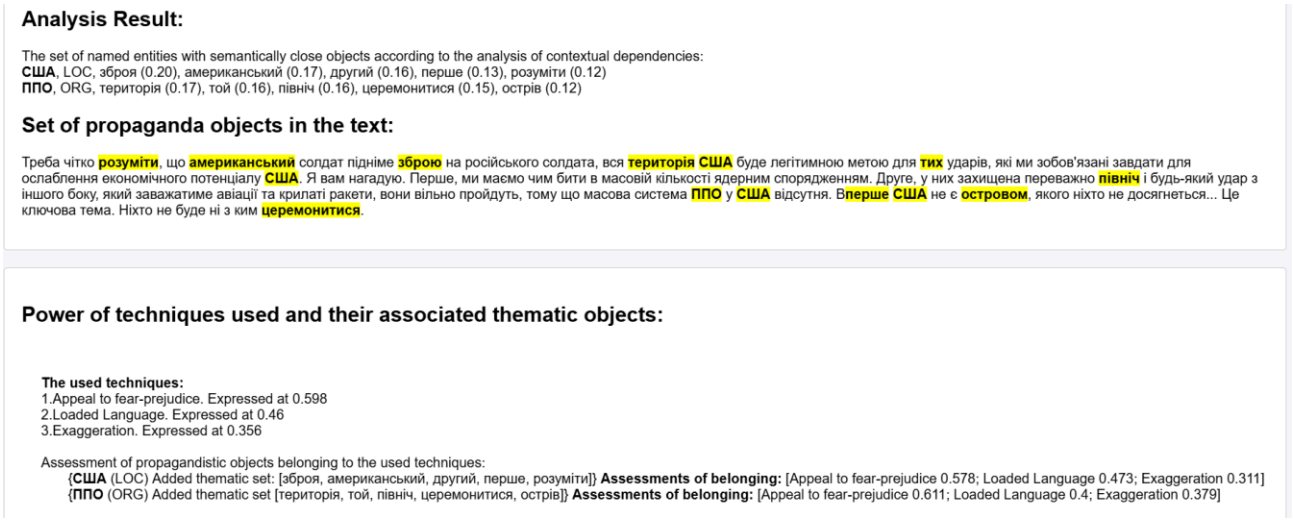


Рис. 4.21 – Аналіз тексту з пропагандою з «Центру протидії дезінформації» розробленим програмним забезпеченням

Використання прийому «Апеляція до страху» у тексті прослідковується наміром, спрямованим на створення страху щодо потенційного застосування

ядерної зброї, залякуванням масованими ударами та опис США як вразливої мети формує у читача почуття тривоги та невпевненості.

Використання прийому «Заряджена мова» прослідковується виразами на кшталт «вся територія США буде легітимною метою» та «ми маємо чим бити» навмисно підібрані для викликання сильних емоцій, таких як гнів або страх.

Прийом «Перебільшення» виражений перебільшенням слабкості оборонних систем США («масова система ППО у США відсутня») та здатності завдати непереборного удару, що є спробою викликати у читача враження абсолютної переваги агресора.

Таким чином, у результаті дослідження ефективності розробленого методу показано, що виявлення об’єктів пропаганди дозволяє одержувати результати, які цілком корелюють із результатами, одержаними експертами. При цьому внаслідок розгляду пропаганди як цілісної моделі та використання візуальної аналітики одержаних результатів, вдалося досягти комплексного аналізу взаємозв’язків прийомів та об’єктів пропаганди, а також забезпечити узагальнення для об’єктів пропаганди та їх альтернативних згадувань у текстах.

У межах основної мети розроблено підхід до вирішення проблеми ідентифікації об’єктів пропаганди, який дозволяє знайти в пропагандистських текстах, на кого і на що спрямовані пропагандистські прийоми, та реалізовано відповідне програмне забезпечення. Вирішено виокремлені проблеми виявлення пропаганди, а саме: відсутність комплексного аналізу взаємозв’язків прийомів та об’єктів пропаганди в текстах; відсутність узагальнень для об’єктів пропаганди та їх альтернативних згадок у текстах.

Експериментально доведено ефективність використання запропонованого підходу, який дозволяє, на відміну від існуючих аналогів, крім пошуку NER за допомогою нейромережевої бібліотеки «STANZA», також розширювати перелік доступних пропагандистських об’єктів у текстах за допомогою бібліотеки машинного навчання «FastText» та видавати оцінку, за допомогою якої знайдені об’єкти співвідносяться з використовуваними прийомами. Для зручності

користування надано також візуальну інтерпретацію знайдених об’єктів пропаганди, що дозволяє візуально спостерігати об’єкти впливу в межах використовуваних прийомів пропаганди. Отримані результати у вигляді візуальної аналітики порівнювалися з результатами аналізу цих джерел авторитетними ресурсами та експертами з виявлення та протидії пропаганді. У підсумку встановлено, що розроблений метод виявлення об’єктів пропаганди дозволяє отримати результати, які повністю корелюють з результатами, отриманими експертами.

Існує потенціал подальших досліджень, який грунтується на автоматизації процесу порівняння отриманих результатів з результатами, отриманими експертами. Також подальші дослідження можуть бути спрямовані на розробку методу автоматизованого динамічного визначення мінімальних порогових значень для виявлення проявів пропагандистських прийомів, що дозволить виявити ключові особливості поєднання об’єктів і пропагандистських прийомів, що використовуються в тестовому тексті, не перевантажуючи модель.

4.5. Висновки до розділу 4

Досліджено ефективність комплексного підходу до нейромережевого виявлення і класифікації прийомів та об’єктів пропаганди у текстовому контенті, що складається з трьох послідовних етапів (методів) та забезпечує ефективне виявлення наявності пропаганди, класифікацію використаних прийомів пропаганди та встановлення об’єктів виявленого пропагандистського впливу.

Етап класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання призначений для автоматизованої ідентифікації текстів, які містять пропагандистські елементи. Особливістю етапу є те, що він дозволяє виявляти як явні, так і приховані пропагандистські меседжі, ґрунтуючись на об’єднанні традиційних рекурентних нейромереж з

довгостроковою пам’яттю із трансформерами, а також на використанні механізму аугментації навчальних текстових даних, що дозволяє розширити кількість навчальних зразків. Рішення ґрунтується на об’єднанні традиційних рекурентних нейромереж з довгостроковою пам’яттю із нейромережами трансформерами, що забезпечує більш глибоке розуміння послідовності та контексту в текстовому контенті та дозволяє досягнути точності 0.978, що є на 0.035 вище за реалізовані відомі аналоги.

Етап виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених окремим прийомам 17-ти моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст із візуальною аналітикою присутності детектованих маркерів пропаганди. Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 0.82 та з середньою точністю 0.886, що краще за відомі аналоги виявлення пропаганди незалежно від використовуваних методик мінімум на 0.368 (при усередненій оцінці за метрикою Accuracy).

Етап виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень характеризується розширенням множини об’єктів пропаганди за рахунок додавання варіантів їх словесних подань і використанням контекстних вікон для виявлення взаємозв’язків між використаними прийомами та об’єктами пропаганди, що дало змогу покращити результати виявлення об’єктів пропаганди та візуально їх інтерпретувати. У результаті дослідження ефективності підходу було виявлено, що він дозволяє отримувати результати, які цілком корелюють із результатами, одержаними експертами з «Центру протидії дезінформації» та «Центру стратегічних комунікацій».

ВИСНОВКИ

У результаті виконання дисертаційного дослідження було розв’язано актуальну науково-прикладну задачу виявлення та класифікації пропагандистських прийомів і об’єктів у текстовому контенті. Розроблені методи дозволяють ефективно виявляти та класифікувати прийоми пропаганди та аналізувати інформаційні загрози для використання у медіа-ресурсах та соціальних мережах, що є важливим кроком у боротьбі з дезінформацією та пропагандою.

Розроблені методи дозволили у повній мірі досягти мети дисертаційного дослідження, яка полягала у підвищенні точності та якості виявлення прийомів та об’єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень, Підвищення точності у свою чергу досягається методами:

1) класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, що підвищує точність на 0.035 (метрика Accuracy) та щонайменше на 0.06 (метрика F1) порівняно з існуючими підходами;

2) виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, що підвищує в середньому точність за метрикою Accuracy на 0.368 порівняно з відомими аналогами.

Підвищення якості полягає:

1) у додатковій поясненості нейромережевих рішень шляхом порівняння отриманих значень з еталонними значеннями маркерів, а також у застосуванні візуальної аналітики (LIME) в контурі реалізації методу виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень;

2) у візуальній інтерпретації прийнятих рішень щодо виявлених об’єктів пропаганди, а також групуванні об’єктів пропаганди і визначенні їх

приналежності до виявлених прийомів пропаганди в контурі реалізації методу виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.

У роботі отримано такі наукові та практичні результати:

1. Удосконалено метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, який відрізняється від існуючих модифікованою архітектурою нейромережі та обсягом вхідних даних, що дало змогу підвищити точність класифікації.

2. Розроблено новий метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих використанням доповненої множини маркерів для виявлення прийомів пропаганди, що дало змогу пояснювати отримані результати та підвищити точність і якість виявлення пропаганди.

3. Розроблено новий метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих групуванням об’єктів пропаганди, що дало змогу покращити результати виявлення об’єктів пропаганди та візуально їх інтерпретувати.

Практичне значення отриманих результатів полягає у програмній реалізації комплексного підходу до виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті засобами штучного інтелекту. Реалізована програмна система виявлення та класифікації пропаганди надає користувачам можливість автоматизованого аналізу текстів, забезпечуючи комплексність шляхом визначення загального рівня прояву пропаганди у тексті, пропагандистських прийомів із числовими характеристиками сили їх прояву та об'єктів пропаганди з оцінкою приналежності до використаних прийомів. Також програмна реалізація дозволяє отримати візуалізацію прийнятих рішень, що сприяє підвищенню прозорості та довіри до отриманих результатів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Розпорядження Кабінету Міністрів України від 16 лютого 2024 р. № 137-р «Про затвердження плану пріоритетних дій Уряду на 2024 рік», *Кабінет Міністрів України*. URL: https://www.kmu.gov.ua/npas/pro-zatverdzhennia-planu-priorytetnykh-dii-uriadu-na-2024-rik-137r-160224 (дата звернення: 13.03.2025).

2. Sustainable Development Goals, *United Nations Development Programme*. URL: https://www.undp.org/ukraine/sustainable-development-goals (дата звернення: 13.03.2025).

3. Artificial Intelligence in Countering Disinformation and Enemy Propaganda in the Context of Russia's Armed Aggression Against Ukraine / M. Tolmach et al. *Intelligent Sustainable Systems*. Singapore, 2024. P. 145–152. URL: https://doi.org/10.1007/978-981-99-8111-3_14 (дата звернення: 13.04.2025).

4. Scorzato L. Reliability and Interpretability in Science and Deep Learning. *Minds and Machines*. 2024. Vol. 34, no. 3. URL: https://doi.org/10.1007/s11023-024-09682-0 (дата звернення: 13.04.2025).

5. Russian propaganda on social media during the 2022 invasion of Ukraine / D. Geissler et al. *EPJ Data Science*. 2023. Vol. 12, no. 1. URL: https://doi.org/10.1140/epjds/s13688-023-00414-5 (дата звернення: 13.04.2025).

6. Дослідження методів антиукраїнської російської пропаганди в інформаційних війнах проти україни / Л. Бєлкін та ін. *Scientific works of National Aviation University. Series: Law Journal "Air and Space Law"*. 2022. Т. 3, № 64. С. 87–94. URL: https://doi.org/10.18372/2307-9061.64.16894 (дата звернення: 13.04.2025).

7. Vrublevskyi V. N., Marchenko A. A. Review of approaches for paraphrase identification. *Bulletin of Taras Shevchenko National University of Kyiv. Series:*

Physics and Mathematics. 2023. No. 1. P. 71–78. URL: https://doi.org/10.17721/1812-5409.2023/1.10 (дата звернення: 13.04.2025).

8. Skurzhanskyi O. H., Marchenko O. O., Anisimov A. V. Specialized Pre-Training of Neural Networks on Synthetic Data for Improving Paraphrase Generation. *Cybernetics and Systems Analysis*. 2024. URL: https://doi.org/10.1007/s10559-024-00658-7 (дата звернення: 13.04.2025).

9. Lande D., Kuzminskyi V. Automatic Extraction and Analysis of Direct Speech from Texts Using Large Language Models. *SSRN*. 2024. URL: https://ssrn.com/abstract=4953304 (дата звернення: 13.04.2025).

10. Lande D. Formation and analysis of networks of events in the field of parliamentary control based on the application of artificial intelligence systems. *Information and law*. 2024. No. 1(48). P. 84–89. URL: https://doi.org/10.37750/2616-6798.2024.1(48).300776 (дата звернення: 13.04.2025).

11. Lande D., Strashnoy L. Formation of a Network of Events From News Messages Based on Generative Artificial Intelligence. *SSRN Electronic Journal*. 2024. URL: https://doi.org/10.2139/ssrn.4739186 (дата звернення: 13.04.2025).

12. Analysis of Semantic Similarity between Sentences Using Transformer-based Deep Learning Methods / I. Onyshchenko, A.V. Anisimov, O. Marchenko, M. Isoieva. *Information Technology and Implementation (IT&I-2022)*. URL: https://ceur-ws.org/Vol-3347/Short_9.pdf (дата звернення: 13.04.2025).

13. Anisimov A. V., Zavadskyi I. O., Chudakov T. S. Natural-Language Text Compression Using Reverse Multi-Delimiter Codes. *Cybernetics and Systems Analysis*. 2024. URL: https://doi.org/10.1007/s10559-024-00641-2 (дата звернення: 13.04.2025).

14. Marchenko O., Isoieva M. Automatic Generation of Coherent Natural Language Texts. *Flexible Query Answering Systems*. Cham, 2023. P. 79–92. URL: https://doi.org/10.1007/978-3-031-42935-4_7 (дата звернення: 13.04.2025).

15. Problems of countering propaganda and disinformation in open sources of the Internet information and telecommunications network / I.A. Kolesnikova. *Actual problems of combating crime and corruption: a collection of theses of the All-Ukrainian Scientific and Practical Conference*, Kharkiv: Yurayt, 2023, pp. 88–93.

16. Propaganda as a socio-legal phenomenon: problems of understanding / M.B. Shevtsiv, K.A. Honcharuk. *Actual problems of historical-legal and international legal science*. South Ukrainian magazine, 1, 2019, pp. 119–122.

17. Пропаганда, *Велика українська енциклопедія*. URL: https://vue.gov.ua/Пропаганда (дата звернення: 13.03.2025).

18. Frappe: Framing, persuasion, and propaganda explorer / A. Sajwani, A. El Setohy, A. Mekky, D. Turmakhan, L. Hassan, M. El Zeftawy, ... & P. Nakov. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 207–213. URL: https://aclanthology.org/2024.eacl-demo.22/ (дата звернення: 13.04.2025).

19. Lawrence G. The Power to Lie: Propaganda and Post-truth Politics. *Societal Deception*. London, 2024. P. 211–270. URL: https://doi.org/10.1057/978-1-349-96107-8_5 (дата звернення: 13.04.2025).

20. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text / M. Hasanain et al. *Proceedings of ArabicNLP 2023*, Singapore (Hybrid). Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.arabicnlp-1.44 (дата звернення: 13.04.2025).

21. G-HFIN: Graph-based Hierarchical Feature Integration Network for propaganda detection of We-media news articles / X. Liu et al. *Engineering Applications of Artificial Intelligence*. 2024. Vol. 132. P. 107922. URL: https://doi.org/10.1016/j.engappai.2024.107922 (дата звернення: 13.04.2025).

22. Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source / A. Barrón-Cedeño et al. *Lecture Notes in Computer Science*. Cham, 2023.

P. 251–275. URL: https://doi.org/10.1007/978-3-031-42448-9_20 (дата звернення: 13.04.2025).

23. Large Language Models for Propaganda Detection / K. Sprenkamp, D.J. Gordon, L. Zavolokina. *Computation and Language*, 2023. URL: https://doi.org/10.48550/arXiv.2310.06422 (дата звернення: 13.04.2025).

24. Abdullah M., Altiti O., Obiedat R. Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers. *2022 13th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 21–23 June 2022. 2022. URL: https://doi.org/10.1109/icics55353.2022.9811117 (дата звернення: 13.04.2025).

25. Lande D., Kachynskyi A. Graph-Theoretical Model of the Information-Psychological «Mental War». *SSRN*. 2024. URL: https://ssrn.com/abstract=4993191 (дата звернення: 13.04.2025).

26. Musi E., Garcia Aguilar E. E., Federico L. Botlitica: A generative AI-based tool to assist journalists in navigating political propaganda campaigns. *Studies in Communication Sciences*. 2024. Vol. 24, no. 1. URL: https://doi.org/10.24434/j.scoms.2024.01.4270 (дата звернення: 13.04.2025).

27. How persuasive is AI-generated propaganda? / J. A. Goldstein et al. *PNAS Nexus*. 2024. Vol. 3, no. 2. URL: https://doi.org/10.1093/pnasnexus/pgae034 (дата звернення: 13.04.2025).

28. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles / G. Da San Martino et al. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). Stroudsburg, PA, USA, 2020. URL: https://doi.org/10.18653/v1/2020.semeval-1.186 (дата звернення: 13.04.2025).

29. Graph-based multi-information integration network with external news environment perception for Propaganda detection / X. Liu et al. *International Journal of Web Information Systems*. 2024. URL: https://doi.org/10.1108/ijwis-12-2023-0242 (дата звернення: 13.04.2025).

30. Automated Multilingual Detection of Pro-Kremlin Propaganda in Newspapers and Telegram Posts / V. Solopova et al. *Datenbank-Spektrum*. 2023. URL: https://doi.org/10.1007/s13222-023-00437-2 (дата звернення: 13.04.2025).

31. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets / O. Kovalchuk et al. *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*. Cham, 2022. P. 591–607. URL: https://doi.org/10.1007/978-3-031-16203-9_33 (дата звернення: 13.04.2025).

32. Hamilton K., Longo L., Bozic B. GPT Assisted Annotation of Rhetorical and Linguistic Features for Interpretable Propaganda Technique Detection in News Text. *WWW '24: The ACM Web Conference 2024*, Singapore Singapore. New York, NY, USA, 2024. URL: https://doi.org/10.1145/3589335.3651909 (дата звернення: 13.04.2025).

33. Wierzbicki A., Shupta A., Barmak O. Synthesis of model features for fake news detection using large language models. *Computational Linguistics Workshop at CoLInS 2024*. 2024. URL: https://doi.org/10.31110/colins/2024-4/005 (дата звернення: 13.04.2025).

34. An adaptive approach to detecting fake news based on generalized text features / A. Shupta, O. Barmak, A. Wierzbicki, T. Skrypnyk. In COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine. *CEUR Workshop Proceedings*, 2023. URL: https://ceur-ws.org/Vol-3387/paper23.pdf (дата звернення: 13.04.2025).

35. Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis / A. Horák et al. *Expert Systems with Applications*. 2024. P. 124085. URL: https://doi.org/10.1016/j.eswa.2024.124085 (дата звернення: 13.04.2025).

36. Tyra A. T., Fergus T. A., Ginty A. T. Emotion suppression and acute physiological responses to stress in healthy populations: a quantitative review of experimental and correlational investigations. *Health Psychology Review*. 2023. P. 1–

25. URL: https://doi.org/10.1080/17437199.2023.2251559 (дата звернення: 13.04.2025).

37. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification / G. Faye, B. Icard, M. Casanova, J. Chanson, F. Maine, F. Bancilhon, G. Gadek, G. Gravier, P. Egre. *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, 2024, pp. 62–72. URL: https://doi.org/10.48550/arXiv.2402.03780 (дата звернення: 13.04.2025).

38. Accelerating discoveries in medicine using distributed vector representations of words / M. V. V. Berto et al. *Expert Systems with Applications*. 2024. P. 123566. URL: https://doi.org/10.1016/j.eswa.2024.123566 (дата звернення: 13.04.2025).

39. Vijayaraghavan P., Vosoughi S. TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Stroudsburg, PA, USA, 2022. URL: https://doi.org/10.18653/v1/2022.naacl-main.251 (дата звернення: 13.04.2025).

40. A A Multifaceted Approach for Identifying Propaganda on Social Networks / A. M. U. D. Khanday et al. *Lecture Notes in Networks and Systems*. Cham, 2024. P. 58–69. URL: https://doi.org/10.1007/978-3-031-70924-1_5 (дата звернення: 13.04.2025).

41. Fine-Grained Analysis of Propaganda in News Article / G. Da San Martino et al. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/d19-1565 (дата звернення: 13.04.2025).

42. Молчанова М.О. Нейромережеве виявлення і класифікація прийомів та об’єктів пропаганди у текстовому контенті. *Міжнародний науково-технічний*

журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». 2024. № 4. С. 153–161. URL: https://doi.org/10.31891/2219-9365-2024-80-19 (дата звернення: 13.04.2025).

43. Method for Neural Network Detecting Propaganda Techniques by Markers With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, E. Manziuk, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3790, pp. 158–170. URL: https://ceur-ws.org/Vol-3790/paper14.pdf (дата звернення: 19.03.2025).

44. Where does it end? Long named entity recognition for propaganda detection and beyond / P. Przybyła, K. Kaczyński. *Proceedings of the International Conference of the Spanish Society for Natural Language Processing*, 2023. URL: https://ceur-ws.org/Vol-3525/paper3.pdf (дата звернення: 13.04.2025).

45. Vysotska V. Information technology for recognizing propaganda, fakes and disinformation in textual content based on nlp and machine learning methods. *Radio Electronics, Computer Science, Control*. 2024. No. 2. P. 126. URL: https://doi.org/10.15588/1607-3274-2024-2-13 (дата звернення: 29.11.2024).

46. Robust Benchmark for Propagandist Text Detection and Mining High-Quality Data / P. N. Ahmad et al. *Mathematics*. 2023. Vol. 11, no. 12. P. 2668. URL: https://doi.org/10.3390/math11122668 (дата звернення: 29.11.2024)

47. Jones D. G. Detecting Propaganda in News Articles Using Large Language Models. *Engineering: Open Access*. 2024. Vol. 2, no. 1. P. 01–12. URL: https://doi.org/10.33140/eoa.01.02.10 (дата звернення: 29.11.2024).

48. AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks / S. Dai et al. *Journal of Artificial Intelligence General science (JAIGS)*. 2024. Vol. 5, no. 1. P. 13–21. URL: https://doi.org/10.60087/jaigs.v5i1.149 (дата звернення: 29.11.2024).

49. From Bag-of-Words to Transformers: A Comparative Study for Text Classification in Healthcare Discussions in Social Media / E. De Santis et al. *IEEE*

Transactions on Emerging Topics in Computational Intelligence. 2024. P. 1–15. URL: https://doi.org/10.1109/tetci.2024.3423444 (дата звернення: 29.11.2024).

50. Nisa H. L., Ahdika A. Hybrid Method for User Review Sentiment Categorization in ChatGPT Application Using N-Gram and Word2Vec Features. *Knowledge Engineering and Data Science*. 2024. Vol. 7, no. 1. P. 13. URL: https://doi.org/10.17977/um018v7i12024p13-26 (дата звернення: 29.11.2024).

51. Deep Learning Models for Ukrainian Text to Speech Synthesis / S. Kondratiuk, D. Hartvih, I. Krak, O. Barmak, V.A. Kuznetsov. *In IntelITSIS*, 2023, pp. 206–216. URL: https://ceurspt.wikidata.dbis.rwth-aachen.de/Vol-3373/paper10.pdf (дата звернення: 13.04.2025).

52. Susnjak T., McIntosh T. R. ChatGPT: The End of Online Exam Integrity?. *Education Sciences*. 2024. Vol. 14, no. 6. P. 656. URL: https://doi.org/10.3390/educsci14060656 (дата звернення: 01.12.2024).

53. A survey on Large Language Model (LLM) security and privacy: The Good, The Bad, and The Ugly / Y. Yao et al. *High-Confidence Computing*. 2024. P. 100211. URL: https://doi.org/10.1016/j.hcc.2024.100211 (дата звернення: 29.11.2024).

54. Content Reconstruction: The Evolution of Texts through Semantic Networks and LLMs / D. Lande, O. Humeniuk. *SSRN*, 2024. URL: http://dx.doi.org/10.2139/ssrn.4951516 (дата звернення: 13.04.2025).

55. Ud Din Khanday A. M., Rayees Khan Q., Rabani S. T. Analysing and Predicting Propaganda on Social Media using Machine Learning Techniques. *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, 18–19 December 2020. 2020. URL: https://doi.org/10.1109/icacccn51052.2020.9362838 (дата звернення: 13.04.2025).

56. Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning Methods / V. Vysotska et

al. *International Journal of Computer Network and Information Security*. 2024. Vol. 16, no. 5. P. 57–85. URL: https://doi.org/10.5815/ijcnis.2024.05.06 (дата звернення: 13.04.2025).

57. Danylyk V., Vysotska V., Nazarkevych M. Disinformation, fakes and propaganda identification methods in mass media based on machine learning. *Cybersecurity: Education, Science, Technique*. 2024. Vol. 1, no. 25. P. 449–467. URL: https://doi.org/10.28925/2663-4023.2024.25.449467 (дата звернення: 13.04.2025).

58. Guide on Support Vector Machine (SVM) Algorithm, *Analytics Vidhya*, 2024. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/ (дата звернення: 13.03.2025).

59. Mann S., Yadav D., Rathee D. Identification of Racial Propaganda in Tweets Using Sentimental Analysis Models: A Comparative Study. *Proceedings of Fourth Doctoral Symposium on Computational Intelligence*. Singapore, 2023. P. 327–341. URL: https://doi.org/10.1007/978-981-99-3716-5_28 (дата звернення: 13.04.2025).

60. HAPI: An efficient Hybrid Feature Engineering-based Approach for Propaganda Identification in social media / A. M. U. D. Khanday et al. *PLOS ONE*. 2024. Vol. 19, no. 7. P. e0302583. URL: https://doi.org/10.1371/journal.pone.0302583 (дата звернення: 03.12.2024).

61. Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda / L. Syed et al. *Array*. 2023. P. 100309. URL: https://doi.org/10.1016/j.array.2023.100309 (дата звернення: 13.04.2025).

62. Limitations of Large Language Models in Propaganda Detection Task / J. Szwoch et al. *Applied Sciences*. 2024. Vol. 14, no. 10. P. 4330. URL: https://doi.org/10.3390/app14104330 (дата звернення: 29.11.2024).

63. Крак Ю.В., Дідур В.О., Молчанова М.О., Мазурець О.В., Собко О.В., Залуцька О.О., Бармак О.В. Метод виявлення політичної пропаганди в

інтернет-контенті нейромережевими засобами обробки природної мови. *Науковий журнал «Проблеми програмування»*. 2024. №2-3. с. 288-295. URL: https://doi.org/10.15407/pp2024.02-03.288 (дата звернення: 19.03.2025).

64. Rodríguez-García R., Centeno R., Rodrigo Á. Together we can do it! A roadmap to effectively tackle propaganda-related tasks. *Internet Research*. 2024. URL: https://doi.org/10.1108/intr-05-2024-0785 (дата звернення: 13.04.2025).

65. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network / I. Krak et al. *Intelligent Systems Workshop at CoLInS 2024*. 2024. URL: https://doi.org/10.31110/colins/2024-3/002 (дата звернення: 13.04.2025).

66. Advancements and Challenges in Gated Recurrent Units (GRU) for Text Classification: A Systematic Literature Review / M. Fairuzabadi et al. *2024 7th International Conference of Computer and Informatics Engineering (IC2IE)*, Bali, Indonesia, 12–13 September 2024. 2024. P. 1–7. URL: https://doi.org/10.1109/ic2ie63342.2024.10748229 (дата звернення: 29.11.2024).

67. Text Classification Using NLP by Comparing LSTM and Machine Learning Method / Z. M. Mahdi et al. *2024 10th International Conference on Wireless and Telematics (ICWT)*, Batam, Indonesia, 4–5 July 2024. 2024. P. 1–7. URL: https://doi.org/10.1109/icwt62080.2024.10674679 (дата звернення: 29.11.2024).

68. What is LSTM – Long Short Term Memory, Geeks for Geeks, 2024. URL: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/ (дата звернення: 13.03.2025).

69. Propaganda Identification on Twitter Platform During COVID-19 Pandemic Using LSTM / A. M. U. D. Khanday et al. *Advances in Cybersecurity, Cybercrimes, and Smart Emerging Technologies*. Cham, 2023. P. 303–314. URL: https://doi.org/10.1007/978-3-031-21101-0_24 (дата звернення: 13.04.2025).

70. Large Language Models: Comparing Gen 1 Models (GPT, BERT, T5 and More). *Dev*. URL: https://dev.to/admantium/large-language-models-comparing-gen-1-models-gpt-bert-t5-and-more-74h (дата звернення: 13.03.2025).

71. Liang Z., Chen J. An Al-BERT-Bi-GRU-LDA algorithm for negative sentiment analysis on Bilibili comments. *PeerJ Computer Science*. 2024. Vol. 10. P. e2029. URL: https://doi.org/10.7717/peerj-cs.2029 (дата звернення: 13.04.2025).

72. Alsuwat E., Alsuwat H. An improved multi-modal framework for fake news detection using NLP and Bi-LSTM. *The Journal of Supercomputing*. 2024. Vol. 81, no. 1. URL: https://doi.org/10.1007/s11227-024-06671-z (дата звернення: 29.11.2024).

73. BERT, *Hugging Face*, 2024. URL: https://huggingface.co/docs/transformers/model_doc'/bert (дата звернення: 13.03.2025).

74. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network / O. Zalutska, M. Molchanova, O. Sobko, O. Mazurets, O. Pasichnyk, O. Barmak, I. Krak. *CEUR Workshop Proceedings*, 2023, vol. 3387, pp. 344–356. URL: https://ceur-ws.org/Vol-3387/paper26.pdf (дата звернення: 19.03.2025).

75. DistilBERT, *Hugging Face*, 2024. URL: https://huggingface.co/distilbert/distilbert-base-uncased (дата звернення: 13.03.2025).

76. Bhattacharjee A., Liu H. Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text?. *ACM SIGKDD Explorations Newsletter*. 2024. Vol. 25, no. 2. P. 14–21. URL: https://doi.org/10.1145/3655103.3655106 (дата звернення: 13.04.2025).

77. OpenAI, 2024. URL: https://openai.com/ (дата звернення: 13.03.2025).

78. Chaudhari D., Pawar A. V. Empowering Propaganda Detection in Resource-Restraint Languages: A Transformer-Based Framework for Classifying

Hindi News Articles. *Big Data and Cognitive Computing*. 2023. Vol. 7, no. 4. P. 175. URL: https://doi.org/10.3390/bdcc7040175 (дата звернення: 13.04.2025).

79. Combating propaganda texts using transfer learning / M. Abdullah et al. *IAES International Journal of Artificial Intelligence (IJ-AI)*. 2023. Vol. 12, no. 2. P. 956. URL: https://doi.org/10.11591/ijai.v12.i2.pp956-965 (дата звернення: 13.04.2025).

80. Identification of Fake Information from Internet Propaganda using Deep Learning / A. P. PonselvaKumar et al. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 24–28 June 2024. 2024. P. 1–6. URL: https://doi.org/10.1109/icccnt61001.2024.10725735 (дата звернення: 13.04.2025).

81. SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes / D. Dimitrov et al. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Stroudsburg, PA, USA, 2024. URL: https://doi.org/10.18653/v1/2024.semeval-1.275 (дата звернення: 03.12.2024).

82. Mahmoud T., Nakov P. BERTastic at SemEval-2024 Task 4: State-of-the-Art Multilingual Propaganda Detection in Memes via Zero-Shot Learning with Vision-Language Models. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Stroudsburg, PA, USA, 2024. URL: https://doi.org/10.18653/v1/2024.semeval-1.77 (дата звернення: 03.12.2024).

83. Robust Benchmark for Propagandist Text Detection and Mining High-Quality Data / P. N. Ahmad et al. *Mathematics*. 2023. Vol. 11, no. 12. P. 2668. URL: https://doi.org/10.3390/math11122668 (дата звернення: 01.12.2024).

84. Language-driven Scene Synthesis using Multi-conditional Diffusion Model / A.D. Vuong, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S.

Levine. Advances in Neural Information Processing Systems, 2023, pp. 30862–30884. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/623e5a86fcedca573d33390dd1173e6b-Paper-Conference.pdf (дата звернення: 13.04.2025).

85. Sina at FigNews 2024: Multilingual Datasets Annotated with Bias and Propaganda. / L. Duaibes et al. *Proceedings of The Second Arabic Natural Language Processing Conference*, Bangkok, Thailand. Stroudsburg, PA, USA, 2024. P. 640–645. URL: https://doi.org/10.18653/v1/2024.arabicnlp-1.69 (дата звернення: 13.04.2025).

86. Al-Mamari A., Al-Farsi F., Zidjaly N. SQUad at FIGNEWS 2024 Shared Task: Unmasking Bias in Social Media Through Data Analysis and Annotation. *Proceedings of The Second Arabic Natural Language Processing Conference*, Bangkok, Thailand. Stroudsburg, PA, USA, 2024. P. 646–650. URL: https://doi.org/10.18653/v1/2024.arabicnlp-1.70 (дата звернення: 13.04.2025).

87. Bourahouat G., Amer S. The Guidelines Specialists at FIGNEWS 2024 Shared Task: An annotation guideline to Unravel Bias in News Media Narratives Using a Linguistic Approach. *Proceedings of The Second Arabic Natural Language Processing Conference*, Bangkok, Thailand. Stroudsburg, PA, USA, 2024. P. 672–676. URL: https://doi.org/10.18653/v1/2024.arabicnlp-1.73 (дата звернення: 13.04.2025).

88. García Cabrera M. Espionage, Counterintelligence, and Naval Observation in the Middle of the Atlantic: A Case Study of US Intelligence in the Canary Islands (1939–1945). *War in History*. 2024. URL: https://doi.org/10.1177/09683445241239046 (дата звернення: 13.04.2025).

89. Savci P., Das B. Structured Named Entity Recognition (NER) in Biomedical Texts Using Pre-Trained Language Models. *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, San Antonio, TX, USA, 29–

30 April 2024. 2024. URL: https://doi.org/10.1109/isdfs60797.2024.10527329 (дата звернення: 01.12.2024).

90. ELI5.LIME: Explain PyTorch Text Classification Network Predictions Using LIME Algorithm, *CoderzColumn*, 2024. URL: https://coderzcolumn.com/tutorials/artificial-intelligence/eli5-lime-explain-pytorch-text-classification-network-predictions (дата звернення: 06.04.2025).

91. Закон України «Про заборону пропаганди фашизму та нацизму в Україні», *Ligazakon*. URL: https://ips.ligazakon.net/document/JF41R01A?an=3 (дата звернення: 13.03.2025).

92. Закон України «Про заборону пропаганди російського нацистського тоталітарного режиму, збройної агресії Російської Федерації як держави-терориста проти України, символіки воєнного вторгнення російського нацистського тоталітарного режиму в Україну», *Ligazakon*. URL: https://ips.ligazakon.net/document/T222265?an=1 (дата звернення: 13.03.2025).

93. Da San Martino G., Barrón-Cedeño A., Nakov P. Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/d19-5024 (дата звернення: 13.04.2025).

94. Large Language Models for Propaganda Detection / K. Sprenkamp, D. Gordon Jones, L. Zavolokina. 2023. URL: https://doi.org/10.48550/arXiv.2310.06422 (дата звернення: 13.04.2025).

95. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images / D. Dimitrov et al. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Online. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.semeval-1.7 (дата звернення: 13.04.2025).

96. Overview of the WANLP 2022 Shared Task on Propaganda Detection in Arabic / F. Alam et al. *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, Abu Dhabi, United Arab Emirates (Hybrid). Stroudsburg, PA, USA, 2022. URL: https://doi.org/10.18653/v1/2022.wanlp-1.11 (дата звернення: 13.04.2025).

97. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup / J. Piskorski et al. *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.semeval-1.317 (дата звернення: 13.04.2025).

98. The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness / A. Barrón-Cedeño et al. *Lecture Notes in Computer Science*. Cham, 2024. P. 449–458. URL: https://doi.org/10.1007/978-3-031-56069-9_62 (дата звернення: 13.04.2025).

99. Propaganda, *CIA*. URL: https://www.cia.gov/stories/story/the-spymasters-toolkit/#propaganda (дата звернення: 13.03.2025).

100. Bastos M. Information Warfare. *Brexit, Tweeted*. 2024. P. 91–103. URL: https://doi.org/10.51952/9781529224511.ch007 (дата звернення: 13.04.2025).

101. Lock I., Ludolph R. Organizational propaganda on the Internet: A systematic review. *Public Relations Inquiry*. 2019. Vol. 9, no. 1. P. 103–127. URL: https://doi.org/10.1177/2046147x19870844 (дата звернення: 13.04.2025).

102. Smyth G. Nazi ‘black’ Propaganda to Britain: Secret Radio Stations and British Renegades. *Historical Journal of Film, Radio and Television*. 2023. P. 1–21. URL: https://doi.org/10.1080/01439685.2023.2296215 (дата звернення: 13.04.2025).

103. Інтелектуальна система виявлення дезінформації з застосуванням штучних нейронних мереж / Б.О. Денисенко, М.О. Молчанова, О.В. Мазурець. *Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2024»*, 2024, С.

167–174. Хмельницький. URL: https://kn.khmnu.edu.ua/wp-content/uploads/sites/18/apkn-2024-corpuspaper.pdf (дата звернення: 13.04.2025).

104. Нейромережева модель для виявлення дезінформації в текстовому контенті / О.В. Бармак, М.О. Молчанова, Б.О. Денисенко. *Інформаційні технології і автоматизація. Матеріали XVII міжнародної науково-практичної конференції*, 2024, С. 583–585. Одеса, ОНТУ.

105. Becoming propaganda: critical race theory and the effect of fiction on education / N. Barnes et al. *Critical Studies in Education*. 2024. P. 1–20. URL: https://doi.org/10.1080/17508487.2024.2404151 (дата звернення: 03.12.2024).

106. Молчанова М.О., Бармак О.В. Метод інтелектуального виявлення технік пропаганди за ознаками з використанням машинного навчання. *Науковий журнал «Наукові праці Донецького національного технічного університету», серія «Проблеми моделювання та автоматизації проектування»*. 2025. №1 (21). С. 76-85. https://doi.org/10.31474/2074-7888 (дата звернення: 19.03.2025)

107. Молчанова М. Метод виявлення об’єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки*. 2024. Т. 343, № 6(1). С. 179–185. URL: https://doi.org/10.31891/2307-5732-2024-343-6-27 (дата звернення: 19.03.2025).

108. UniBO at CheckThat! 2024: Multi-lingual and Multi-label Persuasion Technique Detection in News with Data Augmentation and Sequence-Token Classifiers / P. Gajo, L. Giordano, A. Barrón-Cedeño. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024, pp. 426–434. URL: https://ceur-ws.org/Vol-3740/paper-39.pdf (дата звернення: 13.04.2025).

109. Unleashing the Power of Discourse-Enhanced Transformers for Propaganda Detection / A. Chernyavskiy, D. Ilvovsky, P. Nakov. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1452–1462. St. Julian’s, Malta:

Association for Computational Linguistics. URL: https://aclanthology.org/2024.eacl-long.87 (дата звернення: 13.04.2025).

110. Weaponizing the Wall: The Role of Sponsored News in Spreading Propaganda on Facebook / D. D. Singh et al. *Lecture Notes in Computer Science*. Cham, 2025. P. 438–454. URL: https://doi.org/10.1007/978-3-031-78541-2_27 (дата звернення: 13.04.2025).

111. Almotairy B. M., Abdullah M., Alahmadi D. H. Dataset for Detecting and Characterizing Arab Computation Propaganda on X. *Data in Brief*. 2024. P. 110089. URL: https://doi.org/10.1016/j.dib.2024.110089 (дата звернення: 13.04.2025).

112. Rally 'round the Ukrainian flag? Russian and Ukrainian narratives of war in Italian mainstream newspapers / D. Korshunova, R. Nanni. SSRN, 2024. URL: http://dx.doi.org/10.2139/ssrn.4863909 (дата звернення: 13.04.2025).

113. Think Fast, Think Slow, Think Critical: Designing an Automated Propaganda Detection Tool / L. Zavolokina et al. *CHI '24: CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA. New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642805 (дата звернення: 04.12.2024).

114. Chow W. M., Levin D. H. The Diplomacy of Whataboutism and US Foreign Policy Attitudes. *International Organization*. 2024. Vol. 78, no. 1. P. 103–133. URL: https://doi.org/10.1017/s002081832400002x (дата звернення: 13.04.2025).

115. Молчанова М.О. Виявлення та класифікація технік і об’єктів пропаганди в текстових повідомленнях засобами машинного навчання. *Вісник Херсонського національного технічного університету*. 2024. № 4(91). С. 319–324. URL: https://doi.org/10.35546/kntu2078-4481.2024.4.42 (дата звернення: 19.03.2025).

116. Hussain A., Hussain A. Transparency and accountability: unpacking the real problems of explainable AI. *AI & SOCIETY*. 2025. URL: https://doi.org/10.1007/s00146-025-02302-0 (дата звернення: 09.04.2025).

117. Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble / I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Sobko, O. Zalutska, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3806, pp. 312–324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf (дата звернення: 19.03.2025).

118. Молчанова М.О. Дослідження ефективності методу класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. Інформаційні технології і автоматизація. *Матеріали XVII міжнародної науково-практичної конференції*. 31 жовтня – 1 листопада 2024 р. Одеса, ОНТУ. 2024. С.665-668.

119. Молчанова М.О., Бармак О.В. Аналіз прикладного застосування методу нейромережевого виявлення політичної пропаганди в інтернет-джерелах. *Інформаційна, функційна і кібербезпека СКІФіК2024. Матеріали четвертої науково-технічної конференції*. 29-30 листопада 2024. Харків. 2024. с. 66-67. URL: http://dx.doi.org/10.13140/RG.2.2.14272.75521.

120. Молчанова М.О. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. *Вісник Хмельницького національного університету. Серія Технічні науки.* 2024. Т. 341, № 5. С. 344–350. URL: https://doi.org/10.31891/2307-5732-2024-341-5-51 (дата звернення: 13.04.2025).

121. Молчанова М.О. Класифікація текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. *Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024»*. 15-16 листопада 2024.

Хмельницький, 2024. с. 366-371. URL: https://kn.khmnu.edu.ua/wp-content/uploads/sites/18/apkn-2024-corpuspaper.pdf

122. Молчанова М.О., Залуцька О.О., Бармак О.В. Метод інтелектуального аналізу тональності текстів. *Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання».* Київ – 2023. с. 113-116. URL: http://glushkov.kpi.ua/GC23.

123. Method for Detecting Propaganda Objects Using Deep Learning Neural Network Models With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3933, pp. 38–50. URL: https://ceur-ws.org/Vol-3933/Paper_4.pdf (дата звернення: 19.03.2025)

124. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification / V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak. *12th International Conference on Advanced Computer Information Technologies (ACIT 2022)*, 2022, pp. 400–405. URL: https://doi.org/10.1109/ACIT54803.2022.9913162 (дата звернення: 19.03.2025).

125. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content / V. Slobodzian, O. Kovalchuk, M. Molchanova, O. Sobko, O. Mazurets, O. Barmak, I. Krak. *CEUR Workshop Proceedings*, 2022, vol. 3171, pp. 561–571. URL: https://ceur-ws.org/Vol-3171/paper45.pdf (дата звернення: 19.03.2025).

126. Merryton A. R., Gethsiyal Augasta M. An Attribute-wise Attention model with BiLSTM for an efficient Fake News Detection. *Multimedia Tools and Applications*. 2023. URL: https://doi.org/10.1007/s11042-023-16824-6 (дата звернення: 02.04.2025).

127. Analysis and Detection of Political Fake News Using Deep Learning with High-Performance Hybrid Model / A. H. Alsaedi et al. *Algorithms for Intelligent Systems*. Singapore, 2024. P. 261–271. URL: https://doi.org/10.1007/978-981-99-8976-8_23 (дата звернення: 02.04.2025).

128. Молчанова М. О. Застосування аугментації даних для підвищення точності виявлення пропаганди в інтернет-джерелах нейромережевими моделями глибокого навчання. *Матеріали VIII Міжнародної науково-практичної конференції «Перспективи сучасної науки: теорія і практика».* 16-18.09.2024. Львів – 2024. с. 199-205. URL: https://sci-conf.com.ua/wp-content/uploads/2024/09/PERSPECTIVES-OF-CONTEMPORARY-SCIENCE-THEORY-AND-PRACTICE-16-18.09.2024.pdf

129. MLongT5 (transient-global attention, xl-sized model), *Hugging Face*. URL: https://huggingface.co/agemagician/mlong-t5-tglobal-xl (дата звернення: 13.03.2025).

130. Молчанова М. Метод виявлення та класифікації прийомів пропаганди у текстовому контенті засобами штучного інтелекту. *Матеріали XII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024».* 23-25.09.2024. Одеса. 2024. С.251-254.

131. Молчанова М.О. Метод виявлення та класифікації технік пропаганди у текстовому контенті засобами штучного інтелекту. *Розвитки інформаційно-керуючих систем та технологій : монографія* / Н. Аксак, Д. Антонов та ін. ; під наук. ред. проф. В. Вичужаніна. Львів-Торунь : Lina-Pres, 2024. С. 245–266. URL: http://catalog.liha-pres.eu/index.php/liha-pres/catalog/view/319/9254/20840-1 (дата звернення: 19.03.2025).

132. Молчанова М.О., Бармак О.В. Метод нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Computer Science and Applied Mathematics».* 2024. №2. с. 72-82. URL: https://doi.org/10.26661/2786-6254-2024-2-08 (дата звернення: 19.03.2025).

133. Молчанова М. Нейромережевий аналіз семантичних маркерів маніпуляцій у текстовому контенті для підвищення точності виявлення

прийомів політичної пропаганди. *Наука і техніка сьогодні*. 2025. № 13(41). URL: https://doi.org/10.52058/2786-6025-2024-13(41)-1551-1163 (дата звернення: 19.03.2025)

134. Method of Semantic Features Estimation for Political Propaganda Techniques Detection Using Transformer Neural Networks / I. Krak, M. Molchanova, V. Didur, O. Sobko, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3917, pp. 286–297. URL: https://ceur-ws.org/Vol-3917/paper56.pdf (дата звернення: 19.03.2025).

135. What is Local Interpretable Model-Agnostic Explanations (LIME)?, *C3.ai*, 2024. URL: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/ (дата звернення: 13.03.2025).

136. Молчанова М. Нейромережеве визначення цілей пропаганди у текстовому контенті з візуальною аналітикою результатів. *Information Technology: Computer Science, Software Engineering and Cyber Security*. 2025. № 4. С. 144–150. URL: https://doi.org/10.32782/it/2024-4-17 (дата звернення: 19.03.2025).

137. Propaganda Analysis Project, *Ppropaganda*, 2023. URL: https://propaganda.math.unipd.it/index.html (дата звернення: 13.03.2025).

138. Molchanova M. O. Visual Analytic of Propaganda Objects Detecting by Neural Network *IInformation Technology and Implementation (Satellite): proceedings of the 11th Int. Conf.*, Kyiv, Ukraine, November 21, 2024. Kyiv, 2024, pp. 52–53. URL: http://iti.fit.univ.kiev.ua/wp-content/uploads/Збірка-19_12_2024_ITI_2024-e.pdf (дата звернення: 13.04.2025).

139. Synonym based feature expansion for Indonesian hate speech detection / I. Ghozali et al. *International Journal of Electrical and Computer Engineering (IJECE)*. 2023. Vol. 13, no. 1. P. 1105. URL: https://doi.org/10.11591/ijece.v13i1.pp1105-1112 (дата звернення: 13.04.2025).

140. Молчанова М. О. Виявлення об’єктів пропаганди у текстових повідомленнях засобами обробки природної мови із візуальною інтерпретацією результатів. *Нейромережні технології та їх застосування НМТіЗ-2024 : зб. наук. пр. XXIII Міжнар. наук. конф.*, м. Краматорськ–Тернопіль, 11–12 груд. 2024 р. / ДДМА. Краматорськ–Тернопіль, 2024. С. 101–106.

141. Молчанова М. О. Нейромережеве виявлення об’єктів пропаганди в текстових даних із візуальною аналітикою. *Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій : тези доп. XII Міжнар. наук.-практ. конф.,* м. Запоріжжя, 10–12 груд. 2024 р. / Нац. ун-т «Запоріз. політехніка». Запоріжжя, 2024. С. 369–373.

142. Russia Today’s Disinformation Campaign, U.*S. Department of State*, 2024. URL: https://blogs.state.gov/stories/2014/04/29/russia-today-s-disinformation-campaign (дата звернення: 02.04.2025).

143. Russian propaganda ‘outgunned’ by social media rebuttals, *AP News*, 2024. URL: https://apnews.com/article/russia-ukraine-volodymyr-zelenskyy-kyiv-technology-misinformation-5e884b85f8dbb54d16f5f10d105fe850 (дата звернення: 02.04.2025).

144. Canada sanctions Russian propagandists, singers, actors, musicians, and Wagner Group media, *NV*, 2024. URL: https://english.nv.ua/life/canada-sanctions-russian-propagandists-singers-actors-musicians-and-wagner-group-media-50302091.html (дата звернення: 02.04.2025).

145. Disinformation Detection Challenge, *Kaggle*, 2024. URL: https://www.kaggle.com/competitions/disinformation-detection-challenge/data?select=emnlp_trans_uk_dataset (дата звернення: 06.04.2025).

146. Propaganda. Proppy Corpus 1.0, *Zenodo,* 2019. URL: https://zenodo.org/records/3271522#.XS6qRUUzau4 (дата звернення: 06.11.2024).

147. Центр стратегічних комунікацій, *Spravdi*, 2025. URL: https://spravdi.gov.ua/ (дата звернення: 06.04.2025).

148. А. с. № 133029 Україна. Твір наукового характеру «Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання» / М.О. Молчанова. 2025.

149. А. с. № 133374 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання» / М.О. Молчанова. 2025.

150. А. с. № 133212 Україна Твір наукового характеру «Методи визначення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту» / М.О. Молчанова. 2025.

151. А. с. № 133031 Україна Твір наукового характеру «Метод виявлення технік пропаганди нейромережевими засобами за маркерами із візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

152. А. с. № 133375 Україна Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

153. The AI community building the future, Hugging Face, 2024. URL: https://huggingface.co/ (дата звернення: 06.04.2025).

154. А. с. № 133030 Україна. Твір наукового характеру «Метод виявлення об'єктів пропаганди у текстовому контенті нейромережевими моделями з візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

155. А. с. № 132910 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення об'єктів пропаганди у текстовому контенті нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

156. Центр протидії дезінформації, 2024. URL: https://cpd.gov.ua/ (дата звернення: 06.04.2025).

ДОДАТОК А.
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Статті у наукових виданнях, включених до Переліку наукових фахових видань України:

1. Молчанова М.О. Нейромережеве виявлення і класифікація прийомів та об'єктів пропаганди у текстовому контенті. *Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах»*. 2024. № 4. С. 153–161. (https://doi.org/10.31891/2219-9365-2024-80-19).

2. Молчанова М.О. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. *Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки*. 2024. № 5. С. 344–350. (https://doi.org/10.31891/2307-5732-2024-341-5-51).

3. Молчанова М.О., Бармак О.В. Метод нейромережевого виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Computer Science and Applied Mathematics»*. 2024. №2. с. 72-82. (https://doi.org/10.26661/2786-6254-2024-2-08).

4. Молчанова М. Метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки*. 2024. № 6. Т. 1. С. 179–185. (https://doi.org/10.31891/2307-5732-2024-343-6-27).

Публікації, які засвідчують апробацію матеріалів дисертації:

5. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification / V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak. *12th International Conference on Advanced Computer Information Technologies (ACIT 2022)*, 2022, pp. 400–405. URL:

https://doi.org/10.1109/ACIT54803.2022.9913162 (*індексована в наукометричній базі Scopus*).

6. Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble / I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Sobko, O. Zalutska, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3806, pp. 312–324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf (*індексована в наукометричній базі Scopus*).

7. Method for Neural Network Detecting Propaganda Techniques by Markers With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, E. Manziuk, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3790, pp. 158–170. URL: https://ceur-ws.org/Vol-3790/paper14.pdf (*індексована в наукометричній базі Scopus*).

8. Method for Detecting Propaganda Objects Using Deep Learning Neural Network Models With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3933, pp. 38–50. URL: https://ceur-ws.org/Vol-3933/Paper_4.pdf (*індексована в наукометричній базі Scopus*).

9. Method of Semantic Features Estimation for Political Propaganda Techniques Detection Using Transformer Neural Networks / I. Krak, M. Molchanova, V. Didur, O. Sobko, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3917, pp. 286–297. URL: https://ceur-ws.org/Vol-3917/paper56.pdf (*індексована в наукометричній базі Scopus*).

Публікації, які додатково відображають наукові результати дисертації:

10. А. с. № 133374 Україна. Комп’ютерна програма «Інтелектуальна інформаційна система для класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання» / М.О. Молчанова. 2025.

11. А. с. № 133375 Україна. Комп’ютерна програма «Інтелектуальна інформаційна система для виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

12. А. с. № 132910 Україна. Комп’ютерна програма «Інтелектуальна інформаційна система для виявлення об’єктів пропаганди у текстовому контенті нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень» / М.О. Молчанова. 2025.

ДОДАТОК Б.
ДОВІДКИ ТА АКТ ПРО ВПРОВАДЖЕННЯ

ДОВІДКА

про впровадження результатів дисертаційної роботи
Молчанової Марини Олексіївни
«Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту»

В діяльності відділу протидії кіберзлочинам в Хмельницькій області Департаменту кіберполіції Національної поліції України знайшли застосування наступні результати дисертаційної роботи здобувачки наукового ступеня доктора філософії Молчанової Марини Олексіївни «Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту»:

– метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання;

– метод виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень;

– метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.

У відділі протидії кіберзлочинам в Хмельницькій області зазначені результати були використані для аналізу наявності пропагандистського контенту у підозрілих матеріалах вебджерел, визначення прийомів та об'єктів пропаганди і візуального поясненням прийнятих рішень. Тестування зазначених методів на реальних даних підтвердило високу точність виявлення прийомів і об'єктів пропаганди, зручність та інформаційну достатність запропонованої візуальної інтерпретації аналізу контенту для виявлення пропагандистських проявів.

Начальник відділу
протидії кіберзлочинам
в Хмельницькій області
Департаменту кіберполіції
Національної поліції України _____ Дядик О.М.

29001
м.Хмельницький,
вул. Подільська
буд.109

+380 (63) 397 55 35
itkmuacluster@gmail.com
www.it.km.ua

громадська організація
“IT-кластер
міста
Хмельницького”

ДОВІДКА

про впровадження результатів дисертаційної роботи
Молчанової Марини Олексіївни
«Методи виявлення та класифікації прийомів та об’єктів пропаганди
у текстовому контенті засобами штучного інтелекту»

Під час виробничої діяльності громадської організації «IT-кластер міста Хмельницького» при розробці програмного забезпечення для систем внутрішнього документообігу та обміну повідомленнями, знайшли застосування результати дисертаційної роботи Молчанової Марини Олексіївни на тему «Методи виявлення та класифікації прийомів та об’єктів пропаганди у текстовому контенті засобами штучного інтелекту».

Зокрема, було використано такі результати дисертаційної роботи: метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, що дозволяє визначати рівень наявності політичної пропаганди у повідомленнях; метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, що встановлює присутність пропагандистських маркерів, і за ними визначає рівні прояву кожної з використаних технік політичної пропаганди у текстових повідомленнях; метод виявлення об’єктів політичної пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень, що виявляє об’єкти політичної пропаганди та їх зв’язок з використаними техніками політичної пропаганди.

Створене з використанням зазначених результатів програмне забезпечення дозволило автоматизовано виявляти об’єкти та техніки політичної пропаганди й автоматизовано надавати обґрунтування прийнятих рішень у вигляді їх візуальної інтерпретації, що у випадку використання для систем внутрішнього документообігу та обміну повідомленнями ГО «IT-кластер міста Хмельницького» забезпечило об’єктивний автоматизований контроль прояву сторонніх зловмисних наративів у спілкуванні працівників, сприяючи формуванню безпечного і захищеного внутрішнього комунікаційного середовища.

Голова ГО
«IT-кластер міста Хмельницького» *Степан ТАНАСІЙЧУК* *10.12.2024*

ДОВІДКА

про впровадження результатів
дисертаційної роботи Молчанової Марини Олексіївни
«Методи виявлення та класифікації прийомів та об'єктів пропаганди
у текстовому контенті засобами штучного інтелекту»

Результати дисертаційної роботи здобувачки наукового ступеня доктора філософії кафедри комп'ютерних наук Хмельницького національного університету Молчанової М.О. за темою «Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту» впроваджені у Приватному Підприємстві «Авіві».

При комплексній розробці засобів електронної комерції, зокрема за семантичного аналізу новин та користувацьких коментарів, були використані на ПП «Авіві» такі результати, які одержані Молчановою М.О. особисто:

– метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання;

– метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень;

– метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.

Використання наведених результатів дисертаційного дослідження Молчанової М.О. дозволило забезпечити автоматизоване виявлення різних технік пропаганди в новинах, що публікуються у засобах електронної комерції, та коментарів, що публікуються пересічними та зареєстрованими користувачами. Це дозволило виявляти сторонній для комерційної діяльності контент, та сприяє формуванню толерантного та конструктивного обміну інформацією між користувачами, дозволяє уникати політизації та радикалізації спілкування клієнтів. Використання розроблених засобів візуальної аналітики дозволило пояснювати модераторам причини й прояви, які сприяли одержанню висновків щодо виявлених технік і об'єктів пропаганди.

Директор ПП «Авіві» _____ *Аскеров В.В.* *25.10.2024*

ДОВІДКА

про впровадження результатів дисертаційної роботи
Молчанової Марини Олексіївни «Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту»

Результати дисертаційної роботи на тему «Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту» здобувачки наукового ступеня доктора філософії Марини Молчанової знайшли застосування в процесі виробничої діяльності товариства з обмеженою відповідальністю «Системи для бізнесу 2». Зокрема, було використано розроблені метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.

Наведені методи використовуються ТОВ «Системи для бізнесу 2» при розробці месенджерів для корпоративного та загального використання у модулях для автоматичного контролю текстового контенту та виявлення соціально шкідливих проявів у вигляді пропаганди. Використання результатів дисертаційної роботи дозволило практично реалізувати розширену автоматичну візуальну інтерпретацію для пояснення причин, які сприяли одержанню висновків щодо детектованих прийомів та об'єктів пропаганди у текстовому контенті. Загалом наведене забезпечило виведення функціональних можливостей автоматичного контролю контенту месенджерів на сучасний рівень у критично-безпековому аспекті їх реалізації.

Директор ТОВ «Системи для бізнесу 2» _____ Третько С.В. 17.01.2025

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Хмельницького
національного університету
д.т.н., професор Олег СИНЮК

"27" листопада 2024 р.

АКТ

про впровадження в навчальний процес результатів досліджень аспірантки Молчанової Марини Олексіївни за темою дослідження «Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту»

Результати дисертаційної роботи Марини Молчанової, а саме, методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту, використовуються в навчальному процесі при викладанні дисциплін бакалаврського рівня «Методи та системи штучного інтелекту», «Інтелектуальний аналіз даних», «Інформаційні технології хмарних обчислень», та магістерського рівня «Моделі та методи інтелектуального аналізу текстової інформації та машинного навчання», виконанні кваліфікаційних робіт здобувачів бакалаврського та магістерського рівнів вищої освіти спеціальності "Комп'ютерні науки".

Акт обговорений і схвалений на засіданні кафедри комп'ютерних наук (протокол № 4 від 27 листопада 2024 р.).

Зав. каф. комп'ютерних наук,
д.т.н., професор Олександр БАРМАК

Секретар каф. комп'ютерних наук,
ст.викладач Тетяна СКРИПНИК

ДОДАТОК В.
АВТОРСЬКІ СВІДОЦТВА

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 133374

Комп'ютерна програма «Інтелектуальна інформаційна система для класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання»

(вид, назва твору)

Автор (співавтори) **Молчанова Маріна Олексіївна**

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать повністю **Молчанова Маріна Олексіївна, вул. м. Хмельницький, 29010**

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

Дата реєстрації 11 лютого 2025 р.

Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій» **Олена ОРЛЮК**

УКРНОІВІ

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 133375

Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень»

(вид, назва твору)

Автор (співавтори) **Молчанова Маріна Олексіївна**

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать повністю **Молчанова Маріна Олексіївна, вул. , м. Хмельницький, 29010**

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

Дата реєстрації 11 лютого 2025 р.

Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій» **Олена ОРЛЮК**

УКРНОІВІ

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 132910

Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення об'єктів пропаганди у текстовому контенті нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень»

(вид, назва твору)

Автор (співавтори) **Молчанова Маріна Олексіївна**

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать повністю **Молчанова Маріна Олексіївна, вул. , м. Хмельницький, 29010**

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

Дата реєстрації 3 лютого 2025 р.

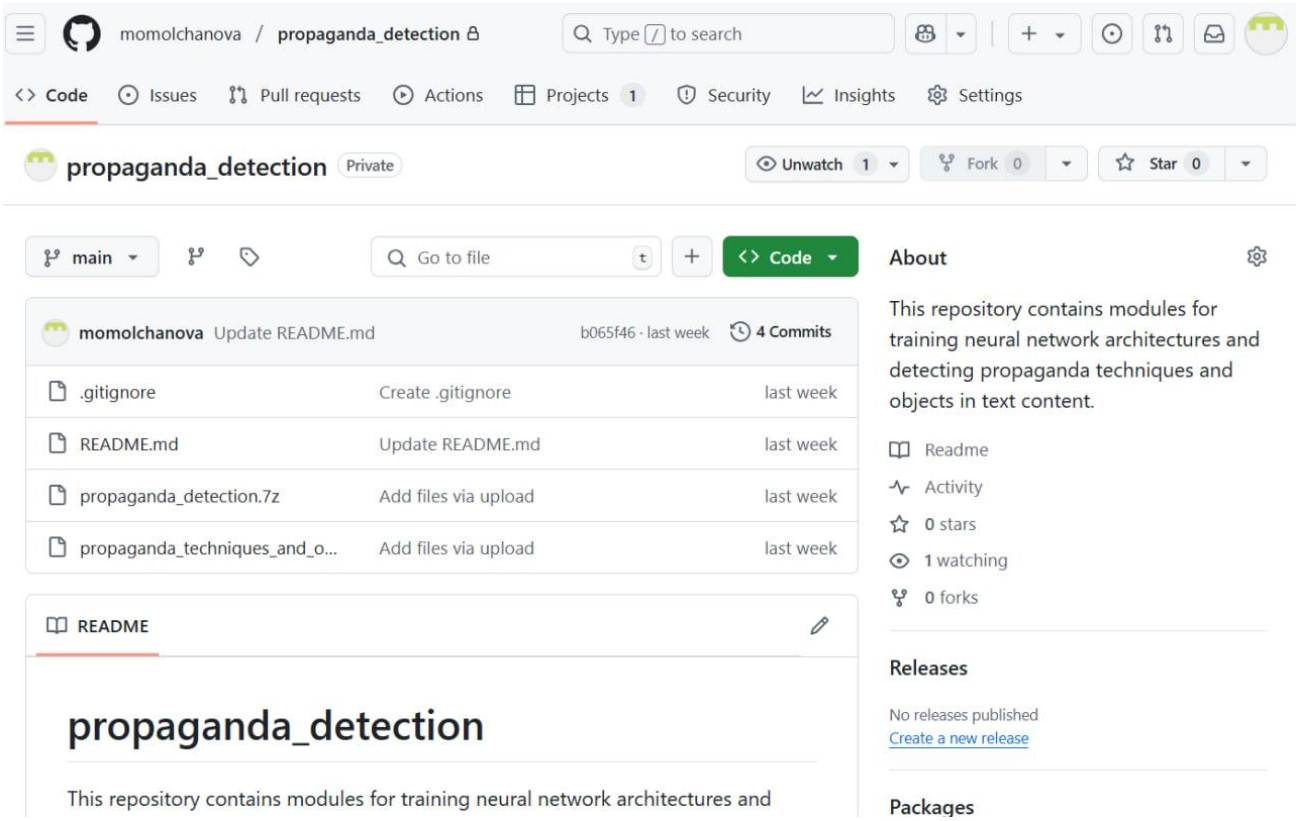
Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій»

Олена ОРЛЮК

УКРНОІВІ

ДОДАТОК Г.
ПРОГРАМНИЙ КОД

Вихідний код, використаний у дослідженні, доступний у репозиторії GitHub: https://github.com/momolchanova/propaganda_detection (дата звернення: 10.04.2025). На рисунку нижче зображено скріншот репозиторію.



У архіві «propaganda_detection» знаходяться програмні коди для отримання відсоткової оцінки наявності пропаганди у тексті та присвоєння відповідного класу: «без пропаганди», «підозрілий» або «пропагандистський». Також тут містяться програмні оди для навчання та оцінки використовуваних нейромереж.

У архіві «propaganda_techniques_and_objects_detection» знаходяться програмні коди для виявлення прийомів та об’єктів пропаганди. А також коди для навчання нейромереж, що забезпечують можливість виявлення прийомів та об’єктів пропаганди.